BIASES OF INCOMPLETE LINEAR MODELS IN FOREST GENETIC DATA ANALYSIS AND OPTIMAL METHODS FOR ESTIMATING TYPE B GENETIC CORRELATIONS

by

PENGXIN LU

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA
1999

ACKNOWLEDGMENTS

I would like to express my gratitude to members of my supervisory committee---Drs. T. L. White, D. A. Huber, R. C. Littell, D. S. Wofford, and D. L. Rockwood ---for their wisdom and efforts in providing invaluable guidance to me throughout my program.

I would especially like to thank Dr. T. L. White, chairman of my supervisory committee, for his support, encouragement, and excellent guidance on many aspects of my study and research, which are crucial to the completion of this project. I am grateful to Dr. T. L. White and the Cooperative Forest Genetics Research Program (CFGRP) at the University of Florida for financial support during the past four years.

I would like to thank Dr. D. A. Huber for his patient guidance, inspiring discussion, and enormous help during the completion of this project.

I am thankful to Mr. Greg Powell for maintaining excellent computing facilities and all his help and friendship.

I would like to thank Dr. G. R. Hodge for his valuable guidance at the early stage of my program.

I extend my thanks to several graduate students in the forest genetics laboratory: Wilson Lopez, Luis Osorio, Jarvia Lopez-Upton, Steve Parker, Lokendra Dhakal, Jeremy Brawner, Victor Sierra, Rebeca Sanhueza, and Ryan Atwood for their friendship and help.

Finally, I would like to thank my parents in China, my wife, Jie Zhang, and my son, Bo Lu, for their understanding, support, and care along the way.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Abstract of Dissertation Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirement for the Degree of Doctor of Philosophy

BIASES OF INCOMPLETE LINEAR MODELS IN FOREST GENETIC DATA ANALYSIS AND OPTIMAL METHODS FOR ESTIMATING TYPE B GENETIC CORRELATIONS

by

Pengxin Lu
May 1999

Chairperson: Timothy L. White
Major Department: School of Forest Resources and Conservation

Potential biases of incomplete mixed models in the estimation of variance component, heritability, and the prediction of breeding gains are theoretically formulated based on balanced data. For a given incomplete mixed model, the magnitudes of biases are functions of population genetic architecture, mating design, and field experimental designs, which can be precisely assessed using the derived formulae. It was found that most incomplete mixed models over-estimate additive genetic variance, resulting in upward-biased heritability and inflated genetic gains. The relative consequence of bias is severe for traits under weak additive genetic control with the strong influence of non-additive genetic effects. For incomplete mixed models ignoring additive genetic effects (GCA) x environment (E) interactions, the potential biases are linearly related to the number of environments included in the data. For incomplete mixed models ignoring dominance effects, biases are linearly

proportional to the number of crosses that each parent is mated. For pure additive genetic models ignoring both dominance effects and GCA x E interaction, the biases are cumulative and can be as high as 60% of the true parameter. For unbalanced data, the formulae can be used to approximate the minimum biases for a given incomplete mixed model by substituting for the average number of design parameters of an experiment.

The search for optimal statistical methods in estimating type B genetic correlations is begun by developing a new univariate approach. The new method estimates type B genetic correlations using predicted parental GCA effects with the technique of best linear unbiased prediction (BLUP) in each individual environment. Numerical comparisons using simulated forest genetic data with various genetic architecture and data imbalance have demonstrated its unbiasedness, better match to underlying true population parameters, and suitability to various experimental designs and data imbalance.

The unbiasedness and precision of multivariate methods in estimating type B genetic correlations are also investigated with a simulation study. It was concluded that constrained multivariate methods produce empirically unbiased estimates of type B genetic correlations which have higher estimation precision, especially when heritabilities of traits are low in the concerned environments. The practical importance of keeping estimates within parameter space and other additional advantages makes the constrained multivariate method a desirable choice.

CHAPTER 1
INTRODUCTION

For genetic evaluation of quantitative traits, mixed model methods have not only become the method of choice in animal breeding (Henderson 1953, 1973, 1984; Mrode 1996), but also gained increasing popularity in tree improvement programs (White and Hodge 1989; Huber 1993; Brralho and Wilson 1994; Jarvis et al. 1995; Dieters et al. 1995; Ericsson and Danell 1995; Wei and Borralho 1998). One of the important applications of mixed linear model theory is the technique of best linear unbiased prediction (BLUP) of breeding values, which has shown superior properties to the traditional approaches of fixed genetic effect models for handling complex messy and unbalanced data structures.

While the theoretical developments of BLUP have been well established (Henderson 1953,1973,1984; Searle et al.1992; Mrode 1996), practical applications of BLUP in breeding programs were hindered by its computational demands (Searle et al. 1992). Although BLUP has now become a routine procedure in animal breeding due to the development of faster computers and more efficient algorithms, its applications in forest genetic evaluation are relatively new (White and Hodge 1989). The lack of computer software suited to the data properties of forest genetic experiments often leads tree breeders to follow those approaches adopted in animal breeding, resulting in the use of incomplete mixed linear models with respect to forest genetic experiments. Although such model specifications may be adequate

for animal genetic data (Varona et al. 1997), the question remains regarding their adequacies in tree breeding value prediction.

The first part of this dissertation places emphasis on examining the effects of incomplete mixed linear models on genetic parameter estimation and parental breeding value prediction in forest genetic data analyses. Knowledge is needed about the potential biases that may result from the use of incomplete mixed linear models because estimates of genetic parameters (such as narrow sense heritability, type B genetic correlation, and the ratio of dominance variance to additive variance) are major considerations in creating a tree breeding strategy (Zobel and Talbert 1984; White et al. 1993), and predicted breeding values are the basis for ranking candidates, making selections, and evaluating genetic progress in a breeding program (White and Hodge 1989; White et al. 1993; Borralho and Dutkowski 1998).

The potential biases from incomplete mixed models in genetic parameter estimation and parental breeding value prediction are most easily evaluated with balanced data. Since estimates of genetic parameters and predicted breeding values are functions of variance components estimated from data samples (Falconer 1989; Mrode 1996), the essential questions of biases can be answered once the effects of incomplete mixed linear models on the estimation of variance components are determined. Due to the property of orthogonality among experimental factors when data are balanced, closed forms of biases for variance component estimates can be derived in terms of design variables of experiments and true variance components of breeding populations.

While the assessment of biases based on balanced data provides useful information on the acceptability of different incomplete mixed linear models in forest genetic data

analysis, it is still unclear about the possible magnitudes of biases from incomplete models when data are highly unbalanced. It is also of interest whether the derived formulas based on balanced data can be used to approximate biases for unbalanced data. These questions are approached by simulation studies considering typical scenarios of forest genetic tests with different mating designs (field experimental design is restricted to the randomized complete block design with single-tree plots), data imbalance, and a range of genetic architectures reported for major forest species (Yeh and Heaman 1987; Adams et al. 1994; Dieters et al. 1995; Li et al. 1996). The 'goodness of fit' for approximated biases based on theoretical formulae is judged by the regression of realized empirical biases from actual unbalanced data analyses on the theoretically approximated biases.

Defined as the genetic correlation for the same trait measured in different environments, type B genetic correlation (Yamada 1962, Burdon 1977) is an important genetic parameter in tree breeding programs. The utilities of type B genetic correlation in forest tree breeding include its applications in the quantitative study of genotype-by-environment interaction (Burdon 1977; Johnson and Burdon 1990; Woolaston et al. 1991; Adams et al. 1994; Dieters et al. 1995; Pswarayi et al. 1997) and its applications in indirect selection (Jiang 1985; White and Hodge 1989; Wu 1993; Johnson 1997). Practical implications of type B genetic correlation in tree breeding include its influence on the determination of breeding zones (Johnson 1997), the deployment of genetically improved materials, and the estimation of genetic gains from indirect selection (White and Hodge 1989).

Statistical methods for estimating type B genetic correlations have been well established for balanced data with univariate approaches (Yamada 1962; Burdon 1977).

Theoretical considerations and empirical results, however, questioned the general utilities of traditional methods when data are highly unbalanced and variances are heterogeneous across environments (Fernando et al.1984). Given the advances of statistical methods and computing software, better approaches are necessary to overcome those inadequacies.

The investigation of optimal estimating methods of type B genetic correlation is begun by developing a new univariate method. An univariate approach is theoretically developed based on predicted parental GCA effect (i.e., one-half of the predicted parental breeding values). The practical applications of the new approach in type B genetic correlation estimation are compared with existing univariate methods using computer simulated data sets with various types of data imbalance and genetic architectures.

Although univariate approaches may produce theoretically unbiased estimates of type B genetic correlations, large sampling errors of variance component estimates for data sets from each individual environment often lead to estimates of type B genetic correlation being out of theoretical parameter space, especially when heritabilities of trait under investigation are low in one or both of the paired environments. For this reason, multivariate methods are exploited to put data sets from different environments into a closed analytical system which will estimate genetic variances and covariances simultaneously and restrain the estimates of type B genetic correlations within the theoretical parameter space. Although restricted maximum likelihood (REML) estimation of type B genetic correlations from the multivariate approach can constrain the estimates within the parameter space, the properties of unbiasedness and minimal variances of the estimates are not theoretically known. These are assessed in this study based on computer simulated data with known population parameters.

CHAPTER 2
POTENTIAL BIASES OF INCOMPLETE LINEAR MODELS IN GENETIC PARAMETER ESTIMATION AND BREEDING VALUE PREDICTION: BALANCED DATA

Introduction

Estimating genetic parameters and predicting breeding values (BVs) are among the primary objectives of forest genetic tests in tree breeding programs (Zobel and Talbert 1984; White 1987, 1996). Estimates of genetic parameters are important considerations in creating breeding strategies (Zobel and Talbert 1984; White et al. 1993) and predicted breeding values are the basis for ranking candidates, making selections and predicting genetic gains (Henderson 1973, 1977, 1984; Falconer 1981; White and Hodge 1989). Proper analysis of data from forest genetic tests is necessary for obtaining the accurate estimates of genetic parameters crucial for maximizing genetic progress in breeding programs.

From quantitative genetics theory, it is well known that the estimation of heritability and the prediction of breeding values are directly affected by the estimation of variance components, especially the estimates of additive genetic variances (Falconer 1981; Mrode 1996). For a given set of experimental data, the estimates of variance components are affected by both statistical procedures (Searle et al. 1992) and analytical linear models being used (Giertych and Van De Sype 1990; Wei and van der Werf 1993). To obtain unbiased estimates of variance components, forest genetic data analyses have traditionally used as

complete a linear model (i.e., full linear models) as possible (Campbell et al. 1986; Bongarten and Hanover 1986; Hodge and White 1992; Adams et al. 1994; Huber et al. 1994; Dieters et al. 1995). These full analytical linear models consider all environmental and genetic effects that may affect the measurements of the trait(s) under investigation, so that maximum information is extracted from the data. However, for experimental data arising from complex genetic structures and field experimental designs, full mixed linear models become increasingly computationally demanding as the number of levels associated with the model term increases.

In recent practical application of mixed linear model methods to forest genetic evaluation, high computational demands associated with large populations, as well as other analytical difficulties, sometime lead to the use of incomplete mixed models with respect to forest genetic experimental designs, in which either non-additive genetic effects or/and genotype-by-environment (G x E) interactions are omitted from the full mixed linear models (Jarvis et al. 1995; Araújo et al. 1996). Although it is known that ignoring these effects can potentially cause biases to the estimation of additive genetic variance, heritability, breeding values and genetic gains (Henderson 1985; Wei and van der Werf 1993; Quinton and Smith 1997), the potential risks of using incomplete mixed linear models in forest genetic data analyses are not well understood. The objectives of this study are: (1) to formulate theoretically the potential biases of estimates of additive genetic variances, heritabilities and predicted breeding values for balanced data resulting from the use of incomplete mixed linear models; and (2) to demonstrate the effects of mating design and field experimental design on these biases.

Methods

Theoretical Formulation of Biases for Balanced Data

It is known that for balanced data, several variance component estimation techniques, including ANOVA based estimators (such as Henderson's type I, type II, and type III) and restricted maximum likelihood estimation (REML), yield identical solutions for variance components for any mixed or random linear model (Searle et al. 1992). While REML solution of variance component estimates requires the assumption of normality and is an iterative procedure, ANOVA based estimators are obtained straightforwardly by: (i) equating the calculated mean squares (or sum of squares) to their pertinent expectations, and (ii) solving the set of linear equations (Henderson 1953; Searle et al. 1992).

In this study, formulae for estimating biases in terms of design variables and variance components were theoretically derived based on the ANOVA approach. Biases of estimates of additive genetic variance, heritability, and predicted parental breeding values were formulated following the statistical relationship between a full and an incomplete mixed linear models for a given experimental data set. This was achieved by first estimating variance components for a given experimental design based on the expected mean squares of a full mixed linear model. The estimates of additive genetic variance components, heritability and the predicted parental breeding values from the full model were thus regarded as the unbiased estimates or predictions. Then, the sum of squares (as well as its associated degrees of freedoms) of ignored effect(s) from the full model was pooled into that of the appropriate remaining effect(s) in the incomplete model. This leads to the reconstruction of

expected mean squares for effects retained in the incomplete mixed linear model and the re-estimation of variance components. Finally, the differences between the full and incomplete models in the estimates of additive genetic variance, heritability, and predicted parental breeding values were defined as the theoretical biases.

Relationship Between Full and Incomplete Linear Models

For analysis of variance with balanced data, detailed rules have been established for calculating sum of squares, partitioning degrees of freedom, and deriving expected mean squares for each effect in a full linear model (Montgomery 1991; Searle et al. 1992). Abiding by these rules, a full linear model for an experiment with half-sib families tested in a randomized complete block design with multiple-tree plots can be written as:

$$Y_{ijkl} = \mu + E_i + B_{ij} + f_k + fe_{ik} + p_{ijk} + e_{ijkl} \qquad 2\text{-}1$$

where

y_{ijkl} is the observation of the l^{th} tree within the k^{th} family in j^{th} block of the i^{th} environment;

μ is the overall mean;

E_i is the fixed effect of the i^{th} environment, i=1,..., t;

B_{ij} is the fixed effect of the j^{th} block within the i^{th} environment, j=1,...,b;

f_k is the random effect of family effect, f_k ~NID(0, σ_{gca}^2), k=1,...,f ;

fe_{ik} is the random effect of family x environment interaction, fe_{ik} ~NID(0, σ_{ge}^2);

p_{ijk} is the random effect of family x block (environment) interaction (*i.e.,* the so-called plot effect), p_{ijk}~NID(0, σ_p^2);

e_{ijkl} is the random effect of l^{th} tree within the k^{th} family in the j^{th} block of the i^{th} environment, e_{ijkl}~NID(0, σ_e^2), l=1,..., n.

Assuming zero covariance among random effects, the appropriate sum of squares, degrees of freedom, and expected mean squares for each effect are obtained in Table 2-1.

When an incomplete model (ignoring one or more effects from the full linear model) is applied to the same experimental data, these established rules, however, become inadequate because they fail to clarify (*i*) how sums of squares and the degrees of freedom associated with ignored effects will be relocated and (*ii*) how the relocation of sums of squares and degrees of freedom of dropped effects may affect the expected mean squares in the incomplete model. For example, if the effect of family x environment interaction (fe_{ik}) is dropped from Eq. 2-1, the incomplete model then becomes:

$$Y_{ijkl} = \mu + E_i + B_{ij} + f_k + p_{ijk} + e_{ijkl} \qquad 2\text{-}2$$

and the correct ANOVA table cannot be obtained using the rules set out for full linear models.

An effective way to determine the relocation of sums of squares in a incomplete linear model is to examine the structures of sums of squares and expected mean squares calculated under the full model. When an effect is assumed unimportant and consequently dropped from a full linear model, it essentially implies that observed variation due to the dropped effect is assumed to be zero. Specifically, in the above incomplete model, when the effect of family x environment interaction (fe_{ik}) is dropped from the full model, it assumes that $\sigma_{fe}^2 = 0$.

Table 2-1. Sum of squares and expected mean squares for the full and incomplete mixed linear models in an analysis of half-sib families tested with a randomized complete block design and multiple-tree plots.

Source of Variation	df	Sum of Squares	Expected Mean Squares
		Full Model ($y_{ijkl}=E_i+B_{ij}+f_k+fe_{ik}+P_{ijk}+e_{ijkl}$)	
Envir.(E)	t-1	$SS_{Envir.}=bfn\sum_{i=1}^{t}(\overline{y_{i...}}-\overline{y_{....}})^2$	
Block(B/E)	t(b-1)	$SS_{B/E}=fn\sum_{i=1}^{t}\sum_{j=1}^{b}(\overline{y_{ij..}}-\overline{y_{i...}})^2$	
Famly(F)	f-1	$SS_F=tbn\sum_{k=1}^{f}(\overline{y_{..k.}}-\overline{y_{....}})^2$	$\sigma_e^2+n\sigma_p^2+bn\sigma_{fe}^2+tbn\sigma_f^2$
F x E	(t-1)(f-1)	$SS_{FxE}=bn\sum_{i=1}^{t}\sum_{k=1}^{f}(\overline{y_{i.k.}}-\overline{y_{i...}}-\overline{y_{..k.}}+\overline{y_{....}})^2$	$\sigma_e^2+n\sigma_p^2+bn\sigma_{fe}^2$
F x B/E	t(b-1)(f-1)	$SS_{FxB/E}=n\sum_{i=1}^{t}\sum_{j=1}^{b}\sum_{k=1}^{f}(\overline{y_{ijk.}}-\overline{y_{ij..}}-\overline{y_{i.k.}}+\overline{y_{i...}})^2$	$\sigma_e^2+n\sigma_p^2$
Residual	tbf(n-1)	$SS_{Res.}=\sum_{i=1}^{t}\sum_{j=1}^{b}\sum_{k=1}^{f}\sum_{l=1}^{n}(y_{ijkl}-\overline{y_{ijk.}})^2$	σ_e^2
Total	tbfn-1	$SS_{total}=\sum_{i=1}^{t}\sum_{j=1}^{b}\sum_{k=1}^{f}\sum_{l=1}^{n}(y_{ijkl}-\overline{y_{....}})^2$	
		Incomplete model ($y_{ijkl}=E_i+B_{ij}+f_k+P_{ijk}+e_{ijkl}$)	
Envir.(E)	t-1	$SS_{Envir.}=bfn\sum_{i=1}^{t}(\overline{y_{i...}}-\overline{y_{....}})^2$	
Block(B/E)	t(b-1)	$SS_{B/E}=fn\sum_{i=1}^{t}\sum_{j=1}^{b}(\overline{y_{ij..}}-\overline{y_{i...}})^2$	
Famly(F)	f-1	$SS_F=tbn\sum_{k=1}^{f}(\overline{y_{..k.}}-\overline{y_{....}})^2$	$\sigma_{e*}^2+n\sigma_{p*}^2+tbn\sigma_{f*}^2$
F x B/E	(tb-1)(f-1)	$SS_{FxB/E}=n\sum_{i=1}^{t}\sum_{j=1}^{b}\sum_{k=1}^{f}(\overline{y_{ijk.}}-\overline{y_{ij..}}-\overline{y_{..k.}}+\overline{y_{....}})^2$	$\sigma_{e*}^2+n\sigma_{p*}^2$
Residual	tbf(n-1)	$SS_{Res.}=\sum_{i=1}^{t}\sum_{j=1}^{b}\sum_{k=1}^{f}\sum_{l=1}^{n}(y_{ijkl}-\overline{y_{ijk.}})^2$	σ_{e*}^2
Total	tbfn-1	$SS_{total}=\sum_{i=1}^{t}\sum_{j=1}^{b}\sum_{k=1}^{f}\sum_{l=1}^{n}(y_{ijkl}-\overline{y_{....}})^2$	

After assuming variance component $\sigma_{ik}^2 = 0$ in the expected mean squares of the full linear model, it is quickly found that family x environment interaction (F x E) would have the same expected mean squares as that of family x block(environment) interaction (i.e., F x B/E). Therefore, when the effect of family x environment interaction is dropped from the full model, its associated sum of squares and degrees of freedom are likely to be pooled into the effect of family x block(environment) interaction which, in this case, is confirmed because

$$SS_{FxB/E} = n \sum_{i=1}^{t} \sum_{j=1}^{b} \sum_{k=1}^{f} (\overline{y_{ijk.}} - \overline{y_{ij..}} - \overline{y_{i.k.}} + \overline{y_{i...}})^2$$

$$= n \sum_{i=1}^{t} \sum_{j=1}^{b} \sum_{k=1}^{f} [(\overline{y_{ijk.}} - \overline{y_{ij..}} - \overline{y_{..k.}} + \overline{y_{....}}) - (\overline{y_{i.k.}} - \overline{y_{i...}} - \overline{y_{..k.}} + \overline{y_{....}})]^2$$

$$= n \sum_{i=1}^{t} \sum_{j=1}^{b} \sum_{k=1}^{f} (\overline{y_{ijk.}} - \overline{y_{ij..}} - \overline{y_{..k.}} + \overline{y_{....}})^2 - bn \sum_{i=1}^{t} \sum_{k=1}^{f} (\overline{y_{i.k.}} - \overline{y_{i...}} - \overline{y_{..k.}} + \overline{y_{....}})^2. \quad 2\text{-}3$$

Clearly, when $SS_{FxE} = bn \sum_{i=1}^{t} \sum_{k=1}^{f} (\overline{y_{i.k.}} - \overline{y_{i...}} - \overline{y_{..k.}} + \overline{y_{....}})^2$ is not deducted in the calculated sum of squares for the family x block (environment) interaction in the incomplete model, the sum of squares for the family x block (environment) interaction in the incomplete model is increased by exactly the same amount as that of the effect of family x environment interaction in the full model.

Now, consider the effect of the relocation of sum of squares of an ignored effect on the expected mean square of the recipient effect. It can be shown that in the above incomplete model, the expected sum of squares for the recipient effect [i.e., family x block(environment) interaction effect] is:

$$E(SS*_{FxB/E}) = E[n \sum_{i=1}^{t} \sum_{j=1}^{b} \sum_{k=1}^{f} (\overline{y_{ijk.}} - \overline{y_{ij..}} - \overline{y_{..k.}} + \overline{y_{....}})^2]$$

$$= (f-1)(tb-1)(\sigma^2_{e*} + n\sigma^2_{p*}) \quad \text{(see Appendix 2-1)} \qquad 2\text{-}4$$

$$\text{Thus, } E(MS*_{FxB/E}) = \frac{1}{(f-1)(tb-1)} E(SS*_{FxB/E}) = \sigma^2_{e*} + n\sigma^2_{p*} \,. \qquad 2\text{-}5$$

Note that $(f-1)(tb-1) = (t-1)(f-1) + t(b-1)(f-1)$, which further suggests that the degrees of freedom associated with family x environment interaction are also pooled into the family x block (environment) interaction in the incomplete model. Comparing the expected mean square in the incomplete model (Eq.2-5) with that from the full model [i.e., $E(MS_{FxB/E}) = \sigma^2_e + n\sigma^2_p$] for the family x block (environment) interaction (Table 2-1), it appears that they are nearly structurally identical. In fact, the variance components in Eq.2-5 are generally different from their counterparts in the full model because the calculated mean square in the incomplete model is not necessarily equal to that from the full model for the family x block(environment) interaction after the pooling of sum of squares and degrees of freedom.

Further, consider the effect of incomplete models on the expected mean square of an effect that is not directly affected by the relocation of sum of squares and degrees of freedom. For example, the sum of squares for the family effect in the incomplete model is calculated in the same way as that in the full model (Table 2-1), but

$$E[SS*_{family}] = E[bnt \sum_{k=1}^{f} (\overline{y_{..k.}} - \overline{y_{....}})^2]$$

$$= (f-1)(\sigma^2_{e*} + n\sigma^2_{p*} + bnt\sigma^2_{f*}) \qquad \text{(see Appendix 2-2),} \qquad 2\text{-}6$$

and $$E(MS*_{family}) = \frac{1}{(f-1)} E(SS*_{family}) = \sigma^2_{e*} + n\sigma^2_{p*} + bnt\sigma^2_{f*}. \qquad 2\text{-}7$$

Comparing Eq. 2-7 with the expected mean square from the full model for the family effect [i.e., $E(MS_{family}) = \sigma^2_e + n\sigma^2_p + bn\sigma^2_{fe} + bnt\sigma^2_f$], the expected mean square from the incomplete model is not exactly the same (see Table 2-1). The variance component for the family x environment interaction has disappeared in the incomplete model and, thus, the expected variance components for other effects are different from their counterparts in the full model. Nevertheless, numerically, the two equations of expected mean squares for the family effect are equal because they are expectations of the same calculated mean square, which is not affected by the elimination of the fe_{ik} term in the incomplete model.

As indicated with these examples, it can be proven that the following relationships are generally held between an incomplete and its full linear models for balanced data: (1) for a given effect dropped from a full linear model, its sum of squares and the associated degrees of freedom are pooled into an effect which, often being either an interaction term or a nested effect, links directly to the dropped effect (i.e., in matrix expression, the incidence matrix of the recipient effect spans the incidence matrix of the dropped effect); (2) for any effect in a incomplete model, its expected mean square resembles that in a full model except that (*i*) the expected variance components are generally different in interpretation and value from their counterparts in the full model and (*ii*) the expected variance component of the dropped effect does not exist; and (3) for an effect whose sum of squares is not affected by the relocation of sums of squares and degrees of freedom in a incomplete model, its calculated mean square

from the incomplete model is numerically equal to that from a full model, however, the expectations of the mean square from the full and incomplete models are structurally different in terms of the variance components involved.

Mating and Field Experimental Designs

Design scenarios considered in this study included half-sib and full-sib progenies tested in randomized complete block (RCB) experimental designs. The mating design for creating full-sib progenies was the half-diallel mating design which have been widely used in tree breeding (Namkoong and Roberds 1974; Zobel and Talbert 1984; Foster and Bridgwater 1986; Burdon and Van Buijtenen 1990; Huber et al. 1992; Li et al. 1996; White 1996). Half-sib progenies are assumed from a polymix mating design, which is also widely used in tree breeding programs (Zobel and Talbert 1984).

Results

Design variables of an experiment are denoted by: p, the number of parents used in a full-sib mating design; k, the average number of crosses per parent; t, the number of testing locations in the field experimental design; and b, the number of blocks within a location. True variance components were denoted by: σ_{gca}^2, one-quarter of additive genetic variance (σ_a^2); σ_{ge}^2, one-quarter of additive genetic effect x environment interaction; σ_{sca}^2, one-quarter of dominance variance; σ_{se}^2, one-quarter of dominance x environment interaction; and σ_e^2, residual variance component. In all cases, Δ stands for bias and BV_f is the full model predicted parental breeding values.

Case Study

Case I. Half-sib families tested with RCB designs and single-tree plots

Suppose that trees of half-sib families are tested over multiple environment with a randomized complete block (RCB) field experimental design and single-tree plots in each environment. A full analytical model is:

$$Y_{ijk} = \mu + E_i + B_{ij} + f_k + fe_{ik} + e_{ijk} \qquad 2\text{-}8$$

where μ, E_i, B_{ij}, f_k and fe_{ik} have the same definition as in Eq.2-1, and e_{ijk} is the random effect of residual, e_{ijk}~NID(0, σ_e^2), where $\sigma_e^2 = \sigma_p^2 + \sigma_e^2$ from 2-1.

By the rules set forth for full models, an ANOVA table is given in Table 2-2.

For the same data, an incomplete model ignoring family x environment interaction is: $Y_{ijk} = \mu + E_i + B_{ij} + f_k + e_{ijk}$, 2-9

which has the same definitions of elements as in Eq.2-8.

By the relationship between full and incomplete linear models for balanced data, it is known that when the effect of family x environment interaction is dropped, its associated sum of squares and degrees of freedom are pooled into the error term. Thus, the estimated variance component for the error term from the incomplete model is

$$\sigma_{e*}^2 = \frac{(t-1)(f-1)(\sigma_e^2 + b\sigma_{ge}^2) + t(b-1)(f-1)\sigma_e^2}{(t-1)(f-1) + t(b-1)(f-1)} = \sigma_e^2 + \frac{b(t-1)\sigma_{ge}^2}{(t-1) + t(b-1)} = \sigma_e^2 + \frac{b(t-1)\sigma_{ge}^2}{tb-1}, \qquad 2\text{-}10$$

which indicates a bias, as compared with the estimate from full model, of the magnitude:

$$\Delta\sigma_{e*}^2 = \frac{b(t-1)\sigma_{ge}^2}{tb-1}. \qquad 2\text{-}11$$

Table 2-2. Sum of squares and expected mean squares for the full and incomplete models in an analysis of half-sib families tested with a randomized complete block design and single-tree plots.

Source of Variation	df	SS	EMS
		Full Model ($y_{ijk}=E_i+B_{ij}+f_k+fe_{ik}+e_{ijk}$)	
Env.(E)	t-1	$SS_{Envir.}=bf\sum_{i=1}^{t}(\overline{y_{i..}}-\overline{y_{...}})^2$	
Block (B/E)	t(b-1)	$SS_{B/E}=f\sum_{i=1}^{t}\sum_{j=1}^{b}(\overline{y_{ij.}}-\overline{y_{i..}})^2$	
Famly (F)	f-1	$SS_F=tb\sum_{k=1}^{f}(\overline{y_{..k}}-\overline{y_{...}})^2$	$\sigma_e^2+b\sigma_{fe}^2+tb\sigma_f^2$
F x E	(t-1)(f-1)	$SS_{FxE}=b\sum_{i=1}^{t}\sum_{k=1}^{f}(\overline{y_{i.k}}-\overline{y_{i..}}-\overline{y_{..k}}+\overline{y_{...}})^2$	$\sigma_e^2+b\sigma_{fe}^2$
residual	t(b-1)(f-1)	$SS_e=\sum_{i=1}^{t}\sum_{j=1}^{b}\sum_{k=1}^{f}(\overline{y_{ijk}}-\overline{y_{ij.}}-\overline{y_{i.k}}+\overline{y_{i..}})^2$	σ_e^2
Total	tbf-1	$SS_{total}=\sum_{i=1}^{t}\sum_{j=1}^{b}\sum_{k=1}^{f}(y_{ijk}-\overline{y_{...}})^2$	
		Incomplete model ($y_{ijk}=E_i+B_{ij}+f_k+e_{ijk}$)	
Env. (E)	t-1	$SS_{Envir.}=bf\sum_{i=1}^{t}(\overline{y_{i..}}-\overline{y_{...}})^2$	
Block (B/E)	t(b-1)	$SS_{B/E}=f\sum_{i=1}^{t}\sum_{j=1}^{b}(\overline{y_{ij.}}-\overline{y_{i..}})^2$	
Famly (F)	f-1	$SS_F=tb\sum_{k=1}^{f}(\overline{y_{..k}}-\overline{y_{...}})^2$	$\sigma_{e*}^2+tb\sigma_{f*}^2$
Residual	(tb-1)(f-1)	$SS_e=\sum_{i=1}^{t}\sum_{j=1}^{b}\sum_{k=1}^{f}(\overline{y_{ijk}}-\overline{y_{ij.}}-\overline{y_{..k}}+\overline{y_{...}})^2$	σ_{e*}^2
Total	tbf-1	$SS_{total}=\sum_{i=1}^{t}\sum_{j=1}^{b}\sum_{k=1}^{f}(y_{ijk}-\overline{y_{...}})^2$	

Note that the numbers of levels for environment, block within environment, family, and trees within a plot are noticed by *t*, *b*, *f*, and *n*, respectively.

From Eq.2-11, it is known that the estimate of error variance is biased unless t=1 and b>1 (i.e., for single site data with RCB design). Eq.2-11 is not defined when t =b=1 since such an experiment is no longer statistically valid.

The calculated mean square related to the family effect is not affected by the relocation of sum of squares and degrees of freedom in the incomplete model, thus

$$\sigma_{e*}^2+tb\sigma_{f*}^2=\sigma_e^2+b\sigma_{ge}^2+tb\sigma_f^2, \quad \text{(see family effects in the full and incomplete}$$

models in Table 2-2)

and $$\sigma_{f*}^2=\sigma_f^2+\frac{\sigma_e^2+b\sigma_{ge}^2-\sigma_{e*}^2}{tb}=\sigma_f^2+\frac{b\sigma_{ge}^2-\frac{b(t-1)}{tb-1}\sigma_{ge}^2}{tb}=\sigma_f^2+\frac{b-1}{tb-1}\sigma_{ge}^2. \qquad 2\text{-}12$$

Therefore, comparing with the estimate of variance component for the family effect from the full model, the bias for the estimate of family variance component from the incomplete model is:

$$\Delta\sigma_f^2=\frac{b-1}{tb-1}\sigma_{ge}^2, \qquad 2\text{-}13$$

which is a function of field experimental design and the true variance component of the dropped effect. Equation 2-13 indicates that for the special case with t=1 and b>1, the estimated family variance component from the incomplete model will be biased upward with the magnitude of σ_{ge}^2. It is also worthwhile noting that with this special experimental design, the sum of the biases equals the variance of the family x environment interaction, *i. e.,*

$$\Delta\sigma_{f*}^2+\Delta\sigma_{e*}^2=\frac{b-1}{tb-1}\sigma_{ge}^2+\frac{t(b-1)}{tb-1}\sigma_{ge}^2=\sigma_{ge}^2. \qquad 2\text{-}14$$

Assuming that variance component of half-sib families estimates 1/4 of total additive genetic variance (Falconer 1981), the bias for heritability estimate from the incomplete

model is: $\Delta h^2=4\Delta\sigma^2_{f*}/(\sigma^2_{f*}+\sigma^2_{e*}) =4(\frac{b-1}{tb-1}\sigma^2_{ge})/(\sigma^2_f+\sigma^2_{ge}+\sigma^2_e)$

$$=4\frac{(b-1)}{tb-1}\frac{\sigma^2_{ge}}{\sigma^2_p}, \qquad 2\text{-}15$$

and the bias for predicted parental breeding values is

$$\Delta_{BV}=[\frac{(b-1)}{(tb-1)}\frac{\sigma^2_{ge}}{\sigma^2_{gca}}]*BV_f. \qquad 2\text{-}16$$

Case II. Half-sib families tested with RCB designs and multiple-tree plots

For this experimental design, the full and incomplete linear models are respectively given in Eq.2-1 and Eq.2-2 and their ANOVA tables are given in Table 2-1. Again, by the relationship between full and incomplete linear models, it is known that in the incomplete model, the sum of squares and the degrees of freedom associated with the family x environment interaction are pooled into the family x block(environment) effect, thus it can be derived that the theoretical estimates of variance components for the error term, family x block (environment)interaction and family effect in the incomplete model are, respectively,

$$\sigma^2_{e*}=\sigma^2_e \ , \qquad 2\text{-}17$$

$$\sigma^2_{p*}=\sigma^2_p+\frac{b(t-1)\sigma^2_{ge}}{tb-1} \ , \qquad 2\text{-}18$$

and $\sigma^2_{f*}=\sigma^2_f+\frac{b-1}{tb-1}\sigma^2_{ge}.$ 2-19

By the same assumption as in Eq.2-15 (*i.e.*, $\sigma^2_A=4\sigma^2_f$), the bias for heritability estimate from the incomplete model for this experimental design is

$$\Delta h^2=4(\frac{b-1}{tb-1}\sigma^2_{ge})/(\sigma^2_f+\sigma^2_{ge}+\sigma^2_p+\sigma^2_e), \qquad 2\text{-}20$$

which is biased if σ^2_{ge} is not zero. Like that in the case I, the bias for predicted parental breeding values is: $\Delta_{BV}=[\frac{(b-1)}{(tb-1)}\frac{\sigma^2_{ge}}{\sigma^2_{gca}}]*BV_f$. 2-21

The estimate of phenotypic variance from the incomplete model with this experimental design is unbiased since $\sigma^2_{phen}=(\sigma^2_f+\sigma^2_{ge}+\sigma^2_p+\sigma^2_e)$.

Case III. Full-sib families tested with RCB designs and single-tree plots

Suppose trees of full-sib families from a half-diallel (or circular mating) mating design are tested in the field with a randomized complete block (RCB) experimental design and single-tree plots. A complete analytical model is (Huber et al. 1992;1994):

$$y_{ijkl} = \mu + E_i + B_{ij} + g_k + g_l + ge_{ik} + ge_{il} + s_{kl} + se_{ikl} + e_{ijkl} \quad 2\text{-}22$$

where y_{ijkl} is the observation of the kl^{th} cross in the j^{th} block of the i^{th} test;

μ is the overall mean;

E_i is the fixed effect of the i^{th} environment;

B_{ij} is the fixed effect of the j^{th} block within the i^{th} environment;

g_k is the random effect of female general combining ability (gca), g_k ~NID(0, σ_{gca}^2);

ge_{ik} is the random effect of female gca x environment interaction, ge_{ik} ~NID(0, σ_{ge}^2);

g_l is the random effect of l^{th} male gca, g_l ~NID(0, σ_{gca}^2);

ge_{il} is the random effect of male gca x environment interaction, ge_{il} ~NID(0, σ_{ge}^2);

s_{kl} is the random effect of specific combining ability (sca) between the k^{th} female and l^{th} male, s_{kl}~NID(0, σ_s^2);

se_{ikl} is the random effect of sca xenvironment interaction, se_{ikl}~NID(0, σ_{se}^2); and

e_{ijkl} is the random effect of kl^{th} cross in j^{th} block within i^{th} location, e_{ijkl}~NID(0, σ_e^2).

Assuming no reciprocal effects of parents and no covariance between any pairs of random variables in the model, an ANOVA table for the full and some incomplete linear models is given in Table 2-3. For this full-sib mating design, four potential incomplete mixed linear models are discussed below.

SCA x environment interactions ignored

In a full model analysis of full-sib data, the effect of SCA x environment interaction generally accounts for a large proportion of the degrees of freedom in the analytical model. Consequently, the computational demands can be substantially reduced if this effect is ignored. Based on the relationship between the full and incomplete mixed linear models, it is known that when SCA x environment interaction is ignored from the full mixed linear model, its degrees of freedom and sum of squares are pooled into the error term. The potential biases to the estimates of variance components, heritability, and predicted parental breeding values with such an incomplete model thus can be calculated as:

$$\Delta_{\sigma^2_{gca}} = \frac{-(p-1)(b-1)(pk-2)}{k(p-2)[p(t-1)(k-2)+t(b-1)(pk-2)]}\sigma^2_{se}, \qquad 2\text{-}23$$

$$\Delta_{h^2} = \frac{4(\sigma^2_{gca}+\Delta\sigma^2_{gca})}{2\sigma^2_{gca}+\sigma^2_{sca}+2\sigma^2_{ge}+\sigma^2_{e}+[bm+(1-m)(\frac{2(p-1)(t-1)}{kt(p-2)}+1)]\sigma^2_{se}}$$

$$-\frac{4\sigma^2_{gca}}{2\sigma^2_{gca}+\sigma^2_{sca}+2\sigma^2_{ge}+\sigma^2_{se}+\sigma^2_{e}}, \qquad 2\text{-}24$$

where $m = \frac{p(t-1)(k-2)}{p(t-1)(k-2)+t(b-1)(pk-2)}$,

and $\Delta_{BV} = \frac{\Delta\sigma^2_{gca}}{\sigma^2_{gca}} * BV_f.$ 2-25

Table 2-3. Expected mean squares for random effects in the full model (a), incomplete model ignoring GCA x E & SCA x E (b), incomplete model ignoring SCA & SCA x E (c), and incomplete model ignoring SCA, GCA x E, and SCA x E (d) in the analysis of full-sib families tested with randomized complete block design, and single-tree plots.

Source	df	MS	EMS
	Full model (a) ($y_{ijkl}=E_i+B_{ij}+g_k+g_l+s_{kl}+ge_{ik}+ge_{il}+se_{ikl}+e_{ijkl}$)		
Env.	t-1	MS_t	
Block	t(b-1)	MS_b	
GCA	p-1	MS_g	$\sigma^2_e+b\sigma^2_{se}+\frac{bk(p-2)}{p-1}\sigma^2_{ge}+tb\sigma^2_{sca}+\frac{tbk(p-2)}{p-1}\sigma^2_{gca}$
SCA	$p(\frac{k}{2}-1)$	MS_s	$\sigma^2_e+b\sigma^2_{se}+tb\sigma^2_{sca}$
GCA x E	$(t-1)(p-1)$	MS_{ge}	$\sigma^2_e+b\sigma^2_{se}+\frac{bk(p-2)}{p-1}\sigma^2_{ge}$
SCA x E	$p(t-1)(\frac{k}{2}-1)$	MS_{se}	$\sigma^2_e+b\sigma^2_{se}$
Error	$t(b-1)(\frac{pk}{2}-1)$	MS_e	σ^2_e
Total	$\frac{tbpk}{2}-1$		
	Incomplete model (c)($y_{ijkl}=E_i+B_{ij}+g_k+g_l+ge_{ik}+ge_{il}+e_{ijkl}$)		
Env.	t-1	MS_t	
Block	t(b-1)	MS_b	
GCA	p-1	MS_g	$\sigma^2_{e''}+\frac{bk(p-2)}{p-1}\sigma^2_{ge''}+\frac{tbk(p-2)}{p-1}\sigma^2_{gca''}$
GCA x E	$p(\frac{k}{2}-1)$	MS_s	$\sigma^2_{e''}+\frac{bk(p-2)}{p-1}\sigma^2_{ge''}$
Error	$t[b(\frac{pk}{2}-1)-p+1]$	$MS_{e''}$	$\sigma^2_{e''}$
Total	$\frac{tbpk}{2}-1$		

Source	df	MS	EMS
	Incomplete model (b) ($y_{ijkl}=E_i+B_{ij}+g_k+g_l+s_{kl}+e_{ijkl}$)		
Env.	t-1	MS_t	
Block	t(b-1)	MS_b	
GCA	p-1	MS_g	$\sigma^2_{e'}+tb\sigma^2_{sca'}+\frac{tbk(p-2)}{p-1}\sigma^2_{gca'}$
SCA	$p(\frac{k}{2}-1)$	MS_s	$\sigma^2_{e'}+tb\sigma^2_{sca'}$
Error	$(tb-1)(\frac{pk}{2}-1)$	$MS_{e'}$	$\sigma^2_{e'}$
Total	$\frac{tbpk}{2}-1$		
	Incomplete model (d)($y_{ijkl}=E_i+B_{ij}+g_k+g_l+e_{ijkl}$)		
Env.	t-1	MS_t	
Block	t(b-1)	MS_b	
GCA	p-1	MS_g	$\sigma^2_{e'''}+\frac{tbk(p-2)}{p-1}\sigma^2_{gca'''}$
Error	$tb(\frac{pk}{2}-1)-p+1$	$MS_{e'''}$	$\sigma^2_{e'''}$
Total	$\frac{tbpk}{2}-1$		

Note: k is the number of crosses per patent.

Apparently, this incomplete linear model results in downward biases to the estimates of additive genetic variance and heritability. Accordingly, the predicted breeding values are more shrunken than those from the full analytical model.

SCA x environment and GCA x environment interactions ignored

The incomplete linear model ignoring SCA x environment and GCA x environment interactions is

$$y_{ijkl} = \mu + E_i + B_{ij} + g_k + g_l + s_{kl} + e_{ijkl}, \qquad 2\text{-}26$$

which has the same definitions of elements as in Eq.2-22. ANOVA table for this incomplete model is given in Table 2-3.

By the relationships between full and incomplete linear models for balanced data, the sums of squares of dropped effects in this incomplete model are pooled into the error term, thus it can be shown that the variance component estimates for the error term, SCA effect and GCA effect , as compared with those from the full model, are, respectively,

$$\sigma^2_{e'} = \sigma^2_e + \frac{b(t-1)}{tb-1}\sigma^2_{se} + \frac{(t-1)bk(p-2)}{(\frac{pk}{2}-1)(tb-1)}\sigma^2_{ge}, \qquad 2\text{-}27$$

and $$\sigma^2_{gca'} = \sigma^2_{gca} + \frac{1}{t}\sigma^2_{ge}. \qquad 2\text{-}28$$

Assuming that full-sib family variance component estimates one-half of total additive genetic variance (Falconer 1981), heritability estimate from this incomplete model is:

$$h^{*2} = \frac{4\sigma^2_{gca'}}{2\sigma^2_{gca'} + \sigma^2_{s'} + \sigma^2_{e'}} = \frac{4(\sigma^2_{gca} + \frac{1}{t}\sigma^2_{ge})}{2\sigma^2_{gca} + \sigma^2_s + \sigma^2_{se} + \sigma^2_e + \frac{1}{t}[2 + \frac{k(t-1)(p-2)}{\frac{pk}{2}-1}]\sigma^2_{ge}}. \qquad 2\text{-}29$$

and the bias for predicted parental breeding value is:

$$\Delta_{BV}=\frac{\Delta\sigma^2_{gca}}{\sigma^2_{gca}} * BV_f. \qquad 2\text{-}30$$

Note that, in this experiment, the estimate of phenotypic variance component (the denominator in Eq.2-29) is also biased as compared with that from the full model (*i.e.*, $2\hat{\sigma}^2_{gca}+2\hat{\sigma}^2_{ge}+\hat{\sigma}^2_{sca}+\hat{\sigma}^2_{se}+\hat{\sigma}^2_{e}$) unless t=1 or $\sigma^2_{ge}=0$.

SCA and SCA x environment interaction ignored

The incomplete model dropping SCA and SCA x environment interaction is

$$y_{ijkl} = \mu + E_i + B_{ij} + g_k + g_l + ge_{ik} + ge_{il} + e_{ijkl}, \qquad 2\text{-}31$$

which has the same definitions of elements as in Eq.2-22 and ANOVA table is given in Table 2-3. Because the sums of squares of SCA and SCA x environment interaction are also pooled into the error term in the incomplete model, it thus can be shown that the estimates of variance components for the error term, GCA x environment interaction and GCA effect are, respectively,

$$\sigma^2_{e'} = \sigma^2_e + \frac{bp(\frac{k}{2}-1)}{b(\frac{pk}{2}-1)-p+1}(\sigma^2_s+\sigma^2_{se}), \qquad 2\text{-}32$$

$$\sigma^2_{ge'} = \sigma^2_{ge} + [\frac{(b-1)(\frac{pk}{2}-1)}{b(\frac{pk}{2}-1)-p+1}]\frac{p-1}{k(p-2)}\sigma^2_{se} - [\frac{p(\frac{k}{2}-1)}{b(\frac{pk}{2}-1)-p+1}]\frac{p-1}{k(p-2)}\sigma^2_s, \qquad 2\text{-}33$$

and $$\sigma^2_{gca'} = \sigma^2_{gca} + \frac{p-1}{k(p-2)}\sigma^2_s. \qquad 2\text{-}34$$

Equation 2-34 indicates that when SCA and SCA x environment interaction are excluded from the incomplete model, bias to the estimate of additive genetic variance component is mostly affected by the number of crosses per parent.

With the same assumptions as in Eq.2-29, the heritability estimate with this incomplete linear model is:

$$h*^2=\frac{4\sigma^2_{gca\,//}}{2\sigma^2_{gca\,//}+2\sigma^2_{ge\,//}+\sigma^2_{e\,//}}=\frac{4(\sigma^2_{gca}+\frac{p-1}{k(p-2)}\sigma^2_s)}{2\sigma^2_{gca}+2\sigma^2_{ge}+\sigma^2_e+m(\sigma^2_s+\sigma^2_{se})} \qquad 2\text{-}35$$

where $m=\frac{1}{b(\frac{kp}{2}-1)-p+1}[\frac{(b-1)(p-1)(pk-2)}{k(p-2)}-bp(\frac{k}{2}-1)]$.

The bias for predicted parental breeding value is

$$\Delta_{BV}=\frac{\Delta\sigma^2_{gca}}{\sigma^2_{gca}}*BV_f=\frac{p-1}{k(p-2)}\frac{\sigma^2_s}{\sigma^2_{gca}}BV_f\,. \qquad 2\text{-}36$$

From Eq.2-36, it is clear that the estimates of additive genetic variance, heritability, and phenotypic variance from this incomplete model are all biased as compared with those from the full model.

Purely additive genetic model

For the purely additive genetic model, SCA, GCA x environment interaction, and SCA x environment interaction are all ignored. This is the most incomplete model that would be used for the analysis of full-sib progeny test data for parental values:

$$y_{ijkl}=\mu+E_i+B_{ij}+g_k+g_l+e_{ijkl}. \qquad 2\text{-}37$$

Equation 2-37 has the same definitions of elements as in Eq. 2-22. Its ANOVA table is also given in Table 2-3.

Since the sums of squares of the dropped effects in this incomplete model are eventually pooled into the sum of squares of error term, it can be derived that the estimates of variance components for the error term and the GCA effect are, respectively,

$$\sigma^2_{e'''}=\sigma^2_e+\frac{b(\frac{tpk}{2}-t-p+1)}{tb(\frac{pk}{2}-1)-p+1}\sigma^2_{se}+\frac{b(t-1)(p-2)k}{tb(\frac{pk}{2}-1)-p+1}\sigma^2_{ge}+\frac{tbp(\frac{k}{2}-1)}{tb(\frac{pk}{2}-1)-p+1}\sigma^2_s \quad . \qquad 2\text{-}38$$

$$\sigma^2_{gca'''}=\sigma^2_{gca}+[\frac{(b-1)(\frac{pk}{2}-1)}{tb(\frac{pk}{2}-1)-p+1}]\frac{p-1}{k(p-2)}\sigma^2_{se}+\frac{b(\frac{pk}{2}-1)-p+1}{tb(\frac{pk}{2}-1)-p+1}\sigma^2_{ge}+[\frac{(tb-1)(\frac{pk}{2}-1)}{tb(\frac{pk}{2}-1)-p+1}]\frac{p-1}{k(p-2)}\sigma^2_s. \qquad 2\text{-}39$$

Thus heritability estimate from this incomplete model is:

$$h*^2=\frac{4\sigma^2_{gca'''}}{2\sigma^2_{gca'''}+\sigma^2_{e'''}}, \qquad 2\text{-}40$$

and the bias for predicted parental breeding values is $\Delta_{BV}=\frac{\Delta\sigma^2_{gca}}{\sigma^2_{gca}}*BV_f$. 2-41

It is obvious that $\sigma^2_{gca'''}$, $\sigma^2_{e'''}$ and $h*^2$ are all biased as compared with those estimates from the full linear model, and so is the estimate of phenotypic variance.

Discussion

Theoretical consideration of biases for estimates of genetic parameters and predicted parental breeding values from incomplete mixed linear models confirms some previous empirical evidence (Wei and van der Werf 1992; Quinton and Smith 1997). As compared with full models, incomplete mixed linear models cause biases in variance component and heritability estimates, which result in biased predicted parental breeding values, as long as

the true variance components of dropped effects are not zero, regardless of the dropped effects being random or fixed effects. The direction of biases to estimates of additive genetic variance and heritability can be downward or upward, depending on the specific progeny test and the incomplete linear models being used. While ignoring the effect of SCA x environment interaction causes slight under-estimation of additive genetic variance and heritability, for all other incomplete linear models the biases are positive, indicating the risk of upward biases in the estimation of genetic parameters if those incomplete linear models are used. These upward biases in additive genetic variance and heritability estimates result in the larger spread (i.e., variance) among predicted parental breeding values, which in turn results in the proportional upward biases of predicted genetic gains.

Population genetic architecture [parameters such as heritability (h^2), type-B genetic correlation ($r_{B)}$ and the ratio dominance variance to additive variance (γ)] would affect the magnitudes of biases to additive genetic variance and heritability estimates because it determines the magnitudes variance components for dropped effects. For example, the theoretical projection for balanced data (Figure 2-1a & c) indicates that when the population true heritability is high and non-additive genetic control is weak (such as genetic architecture 3 in Table 2-4), the relative biases to heritability estimates by incomplete linear models are comparatively small. Whereas, when true heritability is low and non-additive genetic effects are strong (such as genetic architecture 1 and 2 in Table 2-4), the relative biases to heritability estimates by the incomplete models can be substantial (Figure 2-1).

When all G x E interactions are excluded from the incomplete linear model, estimates of additive genetic variance and heritability are mainly affected by the number of

environments in which each family is tested (Figure 2-1a & b). The relative biases to heritability estimates from the incomplete linear models are dramatically increased when the number of environments is fewer than 5, regardless of the population genetic structures and the number of families involved. The extreme example of such an incomplete linear model is the single-site genetic tests, in which the estimate of additive genetic variance is inflated by the whole amount of variance component associated with the additive genetic effect x environment interaction.

For full-sib families tested with RCB designs and single-tree plots, if the SCA effect and its interaction with environment are excluded from a full linear model, biases to the estimates of additive genetic variance and heritability are mainly affected by the average number of crosses that each parent has in the experiment, but they are free of the effect of the number of environments. Biases become increasingly larger when the number of crosses per parent is fewer than 6 (Figure1-1c). When both G x E interactions and SCA effects are dropped from a linear model for full-sib families tested with RCB designs of single-tree plots, biases to the estimates of additive genetic effect and heritability tend to behave in a cumulative manner. Figure-1d indicates that when both the number of environments and the numbers of crosses per parent are small (<4) in an experiment, the percentage bias to heritability estimate from Model V can be as high as 60% if the true heritability is around 0.1.

When incomplete linear models must be used, the most desirable choice seems to be the one that ignores SCA x environment interaction. Several reasons support this preference, including: (1) the degrees of freedom associated with the model term in data analysis are

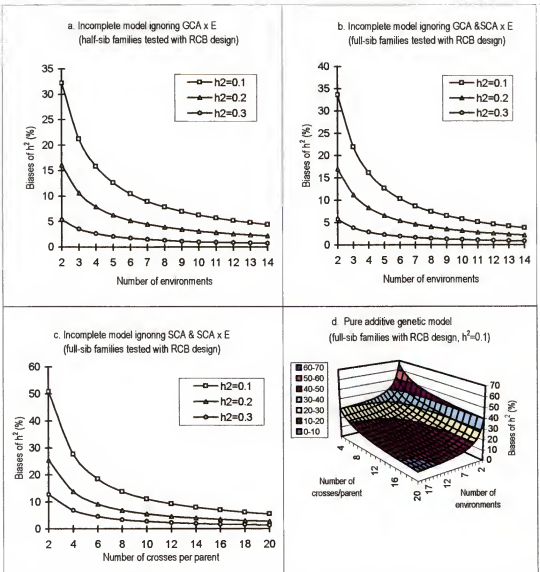

Figure 2-1. Effects of population genetic architecture, mating design, and field experimental design on the biases of estimates of heritability from incomplete linear models. Note: Three levels of genetic architectures are represented by different levels of heritabilites, i.e., for half-sib families, $h^2=0.1$ ($\sigma^2_{gca}=0.25$, $\sigma^2_{ge}=0.1667$, $\sigma^2_{ph}=10$), $h^2=0.2$ ($\sigma^2_{gca}=0.50$, $\sigma^2_{ge}=0.1667$, $\sigma^2_{ph}=10$), and $h^2=0.3$ ($\sigma^2_{gca}=0.75$, $\sigma^2_{ge}=0.0883$, $\sigma^2_{ph}=10$). For full-sib families, $h^2=0.1$ ($\sigma^2_{gca}=0.25$, $\sigma^2_{sca}=0.25$, $\sigma^2_{ge}=0.1667$, $\sigma^2_{se}=0.1667$, $\sigma^2_{ph}=10$), $h^2=0.2$ ($\sigma^2_{gca}=0.50$, $\sigma^2_{sca}=0.25$, $\sigma^2_{ge}=0.1667$, $\sigma^2_{se}=0.0833$, $\sigma^2_{ph}=10$), and $h^2=0.3$ ($\sigma^2_{gca}=0.75$, $\sigma^2_{sca}=0.1875$, $\sigma^2_{ge}=0.0833$, $\sigma^2_{se}=0.0208$, $\sigma^2_{ph}=10$).

reduced dramatically in the incomplete linear models, which greatly reduce computational demands for memory size during data analysis; (2) the magnitude of downward biases for heritability and predicted genetic gains are several-fold smaller than those of biases from dropping other effects; and (3) in reality, the true variance component for SCA x environment interaction is likely to be small.

In addition to biased variance component estimates of additive genetic and other random effects, incomplete linear models also cause biased estimates of phenotypic variances. For the hypothetical cases discussed in this study, although the half-sib families yielded unbiased estimates of phenotypic variance components when G x E interaction was dropped from the full linear model, four incomplete linear models of the full-sib experiments yielded biased estimates of phenotypic variances (as shown in Eq. 2-7b to 2-10b).The direction of biases to the estimates of phenotypic variance can be positive or negative, depending on the specific incomplete linear model being used. However, the magnitude of these biases to phenotypic variance component estimates are small, with negligible effects on the biases to heritability estimates.

Conclusion

ANOVA relationships between full and incomplete linear models for balanced data provide a useful tool to study the mechanisms causing biases in variance component estimation resulting from the use of incomplete mixed linear models. When elements of mating design and field experimental design are known and information about the magnitudes of variance components, such as that for SCA and G x E interaction, is available,

biases to the estimates of additive genetic variance and heritability from incomplete models can be precisely appraised. While the formulae derived from this study are for balanced data, good approximations of biases in the estimates of additive genetic variance, heritability, and predicted parental breeding values may be obtained for unbalanced data by substituting the average values of design variables in the formulae. The 'goodness of fit' for such approximation, however, needs to be examined with different forest genetic scenarios before they can be applied confidently.

CHAPTER 3
POTENTIAL BIASES OF INCOMPLETE LINEAR MODELS IN GENETIC PARAMETER ESTIMATION AND BREEDING VALUE PREDICTION: UNBALANCED DATA

Introduction

Assessment of biases based on balanced data provides useful information on the risk and acceptability of different incomplete mixed linear models in forest genetic data analysis. However, forest genetic testing data are mostly unbalanced due to various reasons (Huber et al. 1994). For both half-sib and full-sib genetic testing data, imbalance can be caused by missing cells, missing observations within a cell or both (White and Hodge 1989; Huber 1993; White 1996). Although closed formulae of biases can be derived in terms of design parameters and variance components for balanced data, no such form, however, can be derived for unbalanced data (Searle et al. 1992). Because unbalanced data structures diminish orthogonality among experimental factors, theoretical formulae of biases derived based on the assumption of orthogonality may become increasingly inaccurate when data are getting increasingly unbalanced. In addition, the more desirable estimation methods of variance components for unbalanced data are those restricted maximum likelihood (REML) approaches rather than the ANOVA based methods (Searle et al. 1992; Huber et al. 1994), which may further increase the discrepancy between the theoretically calculated biases based on balanced data and the real biases from unbalanced data. It is thus unclear how well the

derived formulae based on balanced data can be used to approximate the biases for unbalanced data and uncertainty remains about the potential magnitudes of biases from incomplete models when data are highly unbalanced.

The biases in genetic parameter estimation and breeding value prediction from incomplete linear models are not readily detectable with real experimental data since the true population parameters are never exactly known. Simulated experimental data with known population genetic parameters and different data structures can allow the biases to be precisely investigated under various circumstances, which may in turn provide useful hints to the seriousness of the biases. In this study, we analyze simulated forest genetic testing data of different mating designs, different population genetic architectures (i.e., relative magnitudes of h^2, G x E interaction and dominance variance) and varying data imbalance using the appropriate REML approaches, aiming to: (1) investigate the potential biases of incomplete mixed linear models in estimating heritabilities and predicting of genetic gains for various unbalanced data when dominance effects or/and G x E interactions actually existed in non-inbred populations; (2) to evaluate the feasibility of using those formulae derived based on balanced data to approximate the biases with unbalanced data by using the average design parameters in the experiments.

Methods

Mating and Field Experimental Designs

To avoid the confounding effects of inbreeding on the estimation of additive genetic variance (de Boer and van Arendonk 1992; de Boer and Hoeschele 1993; Hardner et

al.1996), large random mating and non-inbred populations were assumed. Half-diallel, circular and polymix mating designs were used in this study because they are among the most commonly used mating designs in tree breeding (Namkoong and Roberds 1974; Zobel and Talbert 1984; Burdon and van Buijtenen 1990; White et al. 1993). Both half-diallel and circular mating designs are full-sib mating designs and therefore additive and dominance genetic variances as well as G x E interaction variances may be estimated. The polymix mating design represents a half-sib mating design with only estimates of additive genetic variance and G x E interaction variance available. For the half-diallel design, each simulated experiment contained 15 randomly chosen parents from a large parental population, producing 105 crosses (14 crosses per parent). For the circular mating design, 52 randomly chosen parents were included in each experiment with 104 crosses produced (4 crosses per parent, i.e., 1x2, 1x3, 2x3, 2x4, ..., 52x1, 52x2). For the polymix mating design, 105 half-sib families were simulated for each experiment. For each of the mating designs, 500 randomly simulated experiments were investigated for each pertinent level of the genetic architecture and data imbalance.

The field experimental design was maintained as a randomized complete block design with single-tree plots (one individual per family in each block). Families produced in each experiment were assumed to be tested over 4 locations with 15 blocks at each location. Block sizes were restricted to contain no more than 105 trees so that each block can be arranged within 0.1 *ha.* of land under commonly used spacing (Matheson 1989; White 1996). The adoption of single-tree plots was to achieve higher statistical efficiency as recommended by previous studies (Lambeth and Gladstone 1983; Loo-Dinkins and Tauer

1987; Loo-Dinkins et al. 1990; White 1996). Four locations and 15 blocks were used to permit: a) a sufficient number of degrees of freedom to sample G x E interaction variance and b) a reasonable sensitivity to distinguish the real difference between any pairs of families or parents (Cotterill and James 1982).

Data Generation

For experiments with half-diallel and circular mating designs, data were generated from the following mixed linear models by assuming infinitesimal gene action of a continuous trait:

$$y_{ijkl} = \mu + E_i + B_{ij} + g_k + g_l + s_{kl} + ge_{ik} + ge_{il} + se_{ikl} + e_{ijkl}, \qquad 3\text{-}1$$

where y_{ijkl} is the observation on an individual of the kl^{th} cross in the j^{th} block within the i^{th} location;

μ is the overall mean;

E_i is the fixed effect of the i^{th} location;

B_{ij} is the fixed effect of the j^{th} block within the i^{th} location;

g_k is the random effect of the k^{th} female general combining ability (gca), $g_k \sim NID(0, \sigma_{gca}^2)$;

g_l is the random effect of the l^{th} male gca, $g_l \sim NID(0, \sigma_{gca}^2)$;

s_{kl} is the random effect of specific combining ability (sca) between the k^{th} female and the l^{th} male, $s_{kl} \sim NID(0, \sigma_{sca}^2)$;

ge_{ik} is the random effect of interaction between the k^{th} female gca and the i^{th} location, $ge_{ik} \sim NID(0, \sigma_{ge}^2)$;

ge_{il} is the random effect of interaction between the l^{th} male gca and the i^{th} location in,

$ge_{il} \sim NID(0, \sigma_{ge}^2)$;

se_{ikl} is the random effect of interaction between the kl^{th} sca and the i^{th} location, $se_{ikl} \sim NID(0, \sigma_{se}^2)$;

e_{ijkl} is the residual effect of kl^{th} cross (family) in j^{th} block within i^{th} location, $e_{ijkl} \sim NID(0, \sigma_e^2)$.

Reciprocal effects of parents were assumed nil; furthermore, it was assumed that no covariance existed between any pairs of random variables in the model. The effects of location and block were regarded as fixed effects in data analyses, however, for convenience in data generation, the values of location and block effects were taken from normal distributed populations with location $E_i \sim$ NID (0, 20) and block $B_{ij} \sim$NID (0, 10), respectively, across 500 experiments. Since male and female parents are random samples of the same population, which is feasible in mating designs of forest trees (Zobel and Talbert 1984), variances of their GCA effects and the variances of GCA x environment interactions can be pooled, thus

$$E(y_{ijkl}) = \mu + E_i + B_{ij}, \qquad 3\text{-}2$$

$$Var(y_{ijkl}) = 2\sigma_{gca}^2 + 2\sigma_{ge}^2 + \sigma_{sca}^2 + \sigma_{se}^2 + \sigma_e^2. \qquad 3\text{-}3$$

For experiments with polymix mating design, the mixed linear model used for data generation was:

$$y_{ijk} = \mu + E_i + B_{ij} + g_k + ge_{ik} + e_{ijk}, \qquad 3\text{-}4$$

where y_{ijkl} is the observation of the k^{th} family in the j^{th} block of the i^{th} location;

e_{ijk} is the residual effect of k^{th} family in j^{th} block within i^{th} location, $e_{ijk} \sim NID(0, \sigma_e^2)$;

μ, E_i, B_{ij}, g_k and ge_{ik} have the same definition as those in (Eq.3-1).

Again, it was assumed that no covariance between any pair of random variables.

$$E(y_{ijk}) = \mu + E_i + B_{ij}\,, \qquad 3\text{-}5$$

$$Var(y_{ijk}) = \sigma_{gca}^2 + \sigma_{ge}^2 + \sigma_e^2. \qquad 3\text{-}6$$

True Population Genetic Parameters

Population genetic parameters controlling the phenotypic expression of a continuous trait included the narrow sense heritability [i.e., $h^2 = 4\sigma_{gca}^2 / (2\sigma_{gca}^2 + 2\sigma_{ge}^2 + \sigma_{sca}^2 + \sigma_{se}^2 + \sigma_e^2)$ for full-sib and $h^2 = 4\sigma_{gca}^2 / (\sigma_{gca}^2 + \sigma_{ge}^2 + \sigma_e^2)$ for half-sib], type-B genetic correlation [$r_B = \sigma_{gca}^2 / (\sigma_{gca}^2 + \sigma_{ge}^2)$](Yamada 1962; Burdon 1977) and the ratio of dominance variance to additive variance ($\gamma = \sigma_{sca}^2 / \sigma_{gca}^2$). Higher order non-additive genetic effects (such as epistasis) and their interaction with environments were assumed to be absent from the true genetic architectures. Three levels of h^2 (0.1, 0.2 and 0.3), 3 levels of r_B (0.60, 0.75 and 0.90) and 3 levels of γ (0.25, 0.50 and 1.0) were respectively sampled to reflect their potential ranges in forest tree species (Yeh and Heaman 1987; Adams et al. 1995; Dieters et al. 1995; Li et al. 1996). Then, true variance components for each population were consequently derived following Huber et al. (1994) by arbitrarily (but without loss of generality) setting the total phenotypic variance to 10. Instead of considering a factorial combinations of the levels of the three genetic parameters (with 27 different populations), we have chosen three levels of genetic architecture as listed in table 1 to represent low, moderate and high additive genetic control over a phenotypic trait (Table 3-1).

Table 3-1. Genetic parameters and true variance components used to simulate three populations for full-sib and half-sib progeny test data.

Progeny type	Additive genetic control	Genetic parameters			True variance components				
		h^2	γ	r_B	σ_{gca}^2	σ_{sca}^2	σ_{ge}^2	σ_{se}^2	σ_e^2
Full-sib	High	0.30	0.25	0.90	0.7500	0.1875	0.0833	0.0208	8.1251
	Moderate	0.20	0.50	0.75	0.5000	0.2500	0.1667	0.0833	8.3333
	Low	0.10	1.00	0.60	0.2500	0.2500	0.1667	0.1667	8.7500
Half-sib	High	0.30		0.90	0.7500		0.0833		9.1667
	Moderate	0.20		0.75	0.5000		0.1667		9.3333
	Low	0.10		0.60	0.2500		0.1667		9.5833

Data Imbalance

Unbalanced data were simulated by deleting data under each level of the genetic architecture and mating design. For the 15-parent-half-diallel design, three levels of data imbalance were produced, *e.g.,* balanced, moderately unbalanced and severely unbalanced data. Moderately unbalanced data were created by first assuming a random loss of 35 full-sib crosses (1/3 of the total crosses of 105 in an experiment) within each experiment to account for the failure to make crosses or for the complete loss of some full-sib families during the testing period. Then, an average of 40% mortality was assumed to simulate the situations reported in forest genetic tests (Dieters et al. 1995). Severely unbalanced data were created by assuming 70 missing crosses and 40% average mortality. For circular mating design, only two levels of unbalanced data were considered, i.e., the balanced and unbalanced data. The unbalanced data were created by assuming 13 missing crosses (1/8 of the total crosses in a experiment) and 40% average mortality. For polymix mating design, 40% average mortality was assumed for the unbalanced data, but no complete half-sib families were deleted.

In deleting full-sib families from the 15-parent-half-diallel and the 52-parent-circular mating designs, restrictions were applied to ensure that the structures of these mating designs would not be changed, i.e., no genetic disconnectedness would be created among progenies within an experiment. These restrictions implied that at most 13 crosses out of 14 could be deleted for a parent in the 15-parent-half-diallel design, while at most 3 crosses out of 4 could be deleted for a parent in the 52-parent-circular mating design.

While keeping average mortality around 40%, the numbers of missing observations were allowed to fluctuate among families. This implied that at a given site, mortality for some of the families would be much higher than 40% (as high as 90%) while for some other crosses, the mortality could be lower than 40% (as low as 10%). However, for the whole population, average mortality was clustered around 40%. Since mortality was assumed to vary randomly among families across sites, it thus did not contribute extra-binomial genetic variation among families.

Analytical Mixed Linear Models

Five types of mixed linear models were used to analyze each of the simulated data sets. They were: Model I, full models which were identical to those used in data generation (i.e., Eq. 3-1 & 3-4); Model II, SCA x environment interaction excluded from the full model; Model III, all G x E interactions excluded from the full models; Model IV, SCA effect and its interaction with environment excluded from the full model; and Model V, only fixed effects and additive genetic effects considered. The notations and assumptions about these analytical models are the same as those in Eq.3-1 and Eq.3-4.

Variance component estimates and predicted breeding values were obtained using the GAREML computer program (Huber 1993), which uses Giesbrecht's (1983) algorithm to obtain the REML estimation of variance components and then uses these estimates to calculate the best linear unbiased prediction (Henderson 1973) of parental breeding values for a univariate trait. GAREML has repeatedly demonstrated its robustness to starting points and precision in variance component estimation in forest genetic data analysis (Huber 1993; Dieters et al. 1995).

Criteria for Model Evaluation

Heritability

Among the three types of true population genetic parameters (i.e., h^2, r_B and γ), only h^2 could be estimated from all 5 analytical models. Incomplete models ignoring dominance effects (Model II, IV and V) could not estimate γ and incomplete models excluding G x E interaction (Model II, III and V) could not estimate r_B. With the assumptions of no reciprocal effects, no epistasis and no inbreeding, $\hat{\sigma}^2_{gca}$ thus estimated one-quarter of additive genetic variance (Falconer 1981). So, heritability was estimated for each of the 4 analytical models under each of the combinations of mating design, genetic architecture and data imbalance as:

$$\hat{h}^2 = \frac{4\hat{\sigma}^2_{gca}}{\hat{\sigma}^2_p} \qquad 3\text{-}7$$

where $\hat{\sigma}^2_p$ is the total phenotypic variance. For full-sib designs, $\hat{\sigma}^2_p = 2\hat{\sigma}^2_{gca} + 2\hat{\sigma}^2_{ge} + \hat{\sigma}^2_{sca} + \hat{\sigma}^2_{se} + \hat{\sigma}^2_e$ for model I; $\hat{\sigma}^2_p = 2\hat{\sigma}^2_{gca} + 2\hat{\sigma}^2_{ge} + \hat{\sigma}^2_{sca} + \hat{\sigma}^2_e$ for model II, $\hat{\sigma}^2_p = 2\hat{\sigma}^2_{gca} + \hat{\sigma}^2_{sca} + \hat{\sigma}^2_e$ for model III;

$\hat{\sigma}^2_p = 2\hat{\sigma}^2_{gca} + 2\hat{\sigma}^2_{ge} + \hat{\sigma}^2_e$ for model IV; and $\hat{\sigma}^2_p = 2\hat{\sigma}^2_{gca} + \hat{\sigma}^2_e$ for model V. For half-sib designs, $\hat{\sigma}^2_p = \hat{\sigma}^2_{gca} + \hat{\sigma}^2_{ge} + \hat{\sigma}^2_e$ for model I and $\hat{\sigma}^2_p = \hat{\sigma}^2_{gca} + \hat{\sigma}^2_e$ for model III.

For each analytical model under each of the combinations of mating design, genetic architecture and data imbalance, there were 500 estimates of h^2 from 500 simulated experiments. The means of these estimates tended toward convergence. We took the deviations of converged empirical means of 4 analytical models from the true population h^2 as the measures of the magnitudes of biases.

Ratio of predicted genetic gain to true genetic gain (R)

In tree breeding, genetic gain is usually defined as the difference between the mean breeding values of selected individuals from the mean of base population (Falconer 1981; White and Hodge 1989). Since the true breeding values of all parents are known with simulated data, the true genetic gains can be calculated exactly after selection. The ratio of predicted genetic gain to true genetic gain thus shows whether genetic gain is over or under predicted. In this study, it was assumed that 20% of the top ranking parents (*i.e.,* 3, 11 and 21 parents respectively for half-diallel, circular and polymix mating designs) were selected within each experiment (across 4 sites) based on the predicted breeding values. When 5 analytical models were applied to the data of each experiment, there were 5 sets of predicted breeding values and, consequently, 5 sets of selected parents which resulted in 5 pairs of predicted and true genetic gains. For each analytical model, the mean of ratios of predicted genetic gains to true genetic gains over 500 experiments was calculated for a given mating

design, genetic architecture and data imbalance. The deviation of the mean from 1 thus reflected the accuracy of that model in genetic gain prediction.

Reliability of prediction

Reliability is commonly defined as the squared correlation between predicted breeding values and the true breeding values and has been used as a measure of the precision of breeding value predictions (e.g., Uimari and Mäntysaari 1993; Mrode 1996). True estimates of reliability requires the true covariance between predicted and true breeding values as well as their variances. In practice, estimates of reliability are usually obtained based on the estimated covariance from data samples since the true covariance is never known. This implies that the precision of estimated reliability of breeding value predictions can be affected by the precision of estimated genetic covariance and variances. For simulated data with known true BVs, this possibility can be tested by comparing the true and estimated reliabilities for a given data set. The true reliability was simply calculated by its definition in the form:

$$\hat{r}_T^2 = \frac{cov^2(g,\hat{g})}{\hat{\sigma}_T^2 \hat{\sigma}_p^2} \qquad 3\text{-}8$$

where g and $\hat{g}$ are respectively the true and predicted BVs, cov (g, $\hat{g}$) is the covariance between g and $\hat{g}$, and $\hat{\sigma}_T^2$ and $\hat{\sigma}_p^2$ are the estimated variances of g and $\hat{g}$. The estimated reliability was estimated as:

$$\hat{r}_e^2 = 1 - \frac{\hat{\sigma}_{pe}^2}{\hat{\sigma}_{gca}^2} \qquad 3\text{-}9$$

where $\hat{\sigma}_{pe}^2$ is the estimated prediction error variance for gca from the GAREML computer program.

Goodness of Approximation of Biases for Unbalanced Data

The goodness of approximations was evaluated by comparing the realized biases obtained from the actual analyses of unbalanced data with those obtained from the theoretical formulae derived in chapter 2 by using the average design parameters of unbalanced data. Specifically, the simple regression coefficients of approximated biases on the empirically realized biases over repeat random samples of unbalanced data for each of the combinations mating design and genetic architecture was used to reflect the goodness of the approximations.

Results

Empirical Biases

Heritability

Full analytical mixed linear models (Model I), which were identical to those used in data generation, yielded empirically unbiased estimates of heritabilities and were globally superior to any of the incomplete linear models (Table 3-2). Across all mating designs, levels of genetic architecture and degree of data imbalance, none of the mean heritability estimates from the full analytical linear models deviated significantly from the true population parameters. Whereas, for the same sets of data, all incomplete linear models but Model II resulted in significant over estimation of heritability. For the two full-sib mating designs, the most incomplete linear model, ignoring both dominance effects and G x E interaction (*i.e.,* Model V), produced a bias approximately equaling the sum of biases from Model III and IV,

Table 3-2. Mean and empirical standard error of estimates of narrow sense heritability from four different mixed linear models under three mating designs, three levels of genetic architecture, and varying degrees of data imbalance.

Level of genetic control[a]	Model[b]	15-parent-half-diallel			52-parent-circular		105-parent-polymix	
		1[c]	2	3	1	2	1	2
		Mean[d] ± se	Mean ± se	Mean ± se	Mean ± se	Mean ± se	Mean ± se	Mean ± se
1 (low)	I	0.100±0.002	0.101±0.002	0.103±0.003	0.099±0.001	0.098±0.002	0.100±0.001	0.101±0.001
$h^2 = 0.10$	II	0.099±0.002	0.099±0.002	0.099±0.003	0.095±0.001	0.093±0.002	---	---
$r_B = 0.60$	III	0.117±0.002	0.118±0.002	0.120±0.003	0.116±0.001	0.115±0.002	0.116±0.001	0.117±0.001
$\gamma = 1.00$	IV	0.108±0.002	0.113±0.002	0.133±0.003	0.130±0.001	0.136±0.002	---	---
	V	0.126±0.002	0.132±0.002	0.154±0.003	0.151±0.001	0.157±0.002	---	---
2 (moderate)	I	0.195±0.004	0.195±0.004	0.193±0.004	0.199±0.002	0.199±0.002	0.199±0.002	0.200±0.002
$h^2 = 0.20$	II	0.194±0.004	0.193±0.004	0.191±0.004	0.197±0.002	0.196±0.002	---	---
$r_B = 0.75$	III	0.212±0.003	0.213±0.004	0.209±0.004	0.216±0.002	0.216±0.002	0.215±0.002	0.216±0.002
$\gamma = 0.50$	IV	0.202±0.003	0.208±0.004	0.224±0.004	0.230±0.002	0.236±0.002	---	---
	V	0.220±0.003	0.226±0.004	0.242±0.004	0.249±0.002	0.254±0.002	---	---
3 (high)	I	0.297±0.004	0.299±0.004	0.292±0.005	0.300±0.002	0.299±0.003	0.301±0.002	0.300±0.002
$h^2 = 0.30$	II	0.297±0.004	0.299±0.004	0.291±0.005	0.300±0.002	0.298±0.003	---	---
$r_B = 0.90$	III	0.306±0.004	0.307±0.004	0.300±0.005	0.309±0.003	0.306±0.003	0.309±0.002	0.308±0.002
$\gamma = 0.25$	IV	0.303±0.004	0.308±0.004	0.316±0.005	0.323±0.003	0.327±0.003	---	---
	V	0.312±0.004	0.317±0.004	0.325±0.005	0.332±0.003	0.335±0.003	---	---

a. $h^2 = 4\sigma_{gca}^2/\sigma_p^2$; $r_B = \sigma_{gca}^2/(\sigma_{gca}^2 + \sigma_{ge}^2)$; $\gamma = \sigma_{sca}^2/\sigma_{gca}^2$.

b. Types of analytical linear models: I) full model; II) SCA x E interaction excluded, III) SCA x E and GCA x E interactions excluded; IV) SCA and SCA x E interaction excluded, and V) only additive genetic effect considered.

c. Types of data imbalance: for half-diallel design, 1= balanced data, 2=35 missing crosses and 40% mortality, 3= 70 missing crosses and 40% mortality; for circular design, 1=balanced data, 2=13 missing crosses and 40% mortality; for polymix design, 1=balanced data, 2= 40% mortality.

d Each mean is calculated based on estimates of 500 random experiments.

in which either G x E interaction or dominance effect was ignored. Model II, which ignored SCA x environment interaction only, yielded slightly downward biases of heritability estimates but greatly facilitated the procedure of data analysis.

The relative magnitudes of biases in heritability estimates were largest when the population's true heritability and type B genetic correlation were low and the ratio of dominance to additive genetic variances was high. In the genetic architecture level 1 (*i.e.*, h^2=0.1, r_B=0.6 and γ=1.0), 26%-57% relative biases were produced by Model V, which were consistently much higher than those in the genetic architecture level 2 (10%-27%) and level 3 (4%-12%).

The biases in heritability estimates with incomplete linear models were increased by the degree of data imbalance. When a few missing crosses (up to 10 out of 105 crosses) or merely 40% mortality was assumed from the balanced 15-parent-half-diallel mating design, only slight increment of biases were detected for incomplete models (data not shown). However, when more crosses and observations were jointly deleted, the biases in heritability estimates produced by Model IV and V increased dramatically with the decrease in the average number of crosses per parent. In contrast, the biases produced by Model III (omitting G x E interactions) were only slightly affected by data imbalance, which might be the result of holding the field experimental design (i.e., the number of sites and blocks within site) constant across mating designs and data imbalance (Figure 3-1).

Biases of heritability estimates from incomplete models for the circular mating design (balanced and unbalanced data) were comparable to those from the severely unbalanced data

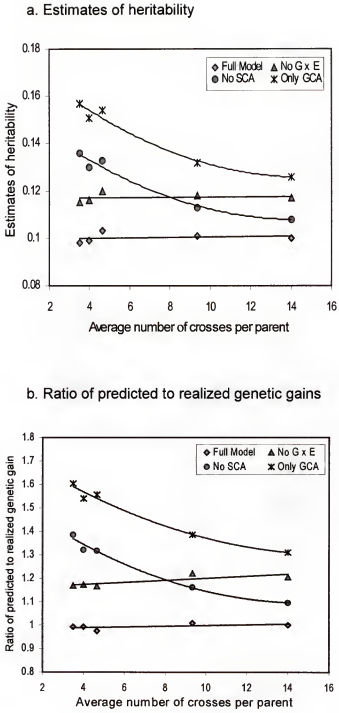

Figure 3-1. Empirical estimates of heritability (a) and the ratio of predicted to realized genetic gains (b) from full and incomplete mixed linear models from different levels of unbalanced data. Data structures are full-sib families created from half-diallel and circular mating designs and tested in a randomized complete block mating design with single-tree plots. True population genetic parameters are: h^2 =0.1, r_B=0.6 and dominance to additive genetic variance component ratio (γ)=1.0.

of the half-diallel mating design. The magnitudes of biases from the incomplete model in polymix mating design were comparable to the biases produced by Model III in the half-diallel and circular mating designs (Table 3-2).

Ratios of predicted to true genetic gains (R)

Similar to heritability estimates, the ratios of predicted genetic gains to true genetic gains were uniformly best for the full analytical mixed linear models (Table 3-3). For the full analytical models, R values were consistently close to 1 for any of the mating designs, genetic architectures and levels of data imbalance. On the other hand, for all incomplete models but Model II (i.e., Model III, IV, and V), R values were significantly larger than 1, which indicated the over-prediction of genetic gains when using predicted parental breeding values from these models. Again, mixed models in which only additive genetic effects were considered have consistently yielded the largest over-prediction (up to 60%), depending on the nature of genetic architecture and the degree of data imbalance of the data. Predicted genetic gain from Model II was, however, slightly lower but not much than from the full model.

The effect of genetic architecture on R showed the same trend as that on heritability. When true population genetic parameters of h^2 and r_B were low but γ was high, larger biases in genetic gain prediction were produced. With strong additive genetic control, (h^2=0.3, r_B=0.90 and γ=0.25), only about 5%-14% over-prediction of genetic gain would occur for any of the incomplete analytical models. This contrasted sharply with the biases obtained in genetic architecture level 1 (weak additive genetic control), in which 10-32% over-prediction of genetic gain would be expected for balanced data in the 15-parent-half-diallel mating

Table 3-3. Mean and empirical standard error of ratios of predicted genetic gains to true genetic gains from four different mixed linear models for three mating designs, three levels of genetic architecture, and varying data imbalance if the top 20% parents were selected in each experiment based on the predicted breeding values.

Level of genetic control[a]	Model[b]	15-parent-half-diallel			52-parent-circular		105-parent Polymix	
		1[c]	2	3	1	2	1	2
		Mean[d] ± se	Mean ± se	Mean ± se	Mean ± se	Mean ± se	Mean ± se	Mean ± se
1 (low)	I	1.001±0.013	1.009±0.018	0.975±0.024	0.993±0.012	0.993±0.015	1.007±0.008	1.009±0.012
$h^2 = 0.10$	II	0.988±0.013	0.990±0.018	0.940±0.024	0.951±0.012	0.949±0.015	---	---
$r_B = 0.60$	III	1.206±0.014	1.222±0.019	1.167±0.024	1.174±0.011	1.169±0.015	1.176±0.008	1.182±0.011
$\gamma = 1.00$	IV	1.097±0.013	1.163±0.019	1.319±0.025	1.323±0.011	1.387±0.015	---	---
	V	1.312±0.015	1.387±0.020	1.557±0.025	1.541±0.012	1.604±0.016	---	---
2 (moderate)	I	1.002±0.010	0.998±0.010	1.007±0.016	1.000±0.007	1.008±0.009	0.995±0.005	0.996±0.006
$h^2 = 0.20$	II	0.999±0.010	0.993±0.010	0.996±0.016	0.988±0.007	0.992±0.009	---	---
$r_B = 0.75$	III	1.106±0.010	1.098±0.010	1.113±0.016	1.089±0.007	1.094±0.009	1.086±0.005	1.078±0.015
$\gamma = 0.50$	IV	1.049±0.010	1.067±0.010	1.197±0.016	1.116±0.007	1.205±0.009	---	---
	V	1.155±0.010	1.177±0.012	1.312±0.018	1.262±0.007	1.300±0.009	---	---
3 (high)	I	0.994±0.005	0.999±0.007	1.003±0.011	1.005±0.005	1.006±0.006	1.004±0.003	1.006±0.005
$h^2 = 0.30$	II	0.994±0.005	0.998±0.007	1.001±0.011	1.003±0.005	1.003±0.006	---	---
$r_B = 0.90$	III	1.026±0.005	1.030±0.007	1.034±0.011	1.029±0.004	1.033±0.006	1.031±0.003	1.034±0.005
$\gamma = 0.25$	IV	1.016±0.005	1.035±0.007	1.099±0.011	1.082±0.004	1.106±0.006	---	---
	V	1.048±0.005	1.065±0.007	1.135±0.011	1.110±0.004	1.137±0.006	---	---

Notations for a, b, and c are the same as those in table 3-2.

d. Means are calculated based on 500 random samples except for unbalanced data set 3 of half-diallel design, in which means are based on 466 to 500 samples due to the deletion of some leverage points.

design and up to 60% over-prediction for balanced and unbalanced data in the 52-parent-circular mating design.

Data imbalance also greatly affected the biases in genetic gain prediction (Table 3-3). Compared to the balanced data in the 15-parent-half-diallel mating design, biases from Model IV and V steadily increased with deletion of more crosses. But, this was not the case for Model III, for which the biases stayed at roughly the same level across all degrees of data imbalance and mating designs even though the average number of crosses and progenies per parent changed considerably. Balanced data of the 52-parent-circular mating design had similar average number of crosses per parent as that in the most severely unbalanced case of 15-parent-half-diallel mating design, and they yielded similar magnitudes of biases from the incomplete models.

Despite the over or slightly under prediction of genetic gains, incomplete models achieved almost the same amount of true genetic gain as that achieved by full analytical model for a given data set with the same selection method (data not shown). Although true genetic gains from selection were significantly affected by the combination of genetic architecture, mating design and data imbalance, little difference in true genetic gain was produced by the five different analytical models for any of the data sets. Whereas, predicted genetic gains depended on the exact values of the predicted breeding values and were greatly influenced by the incomplete linear models. The true genetic gain achieved by selecting the top parents depended only on the rankings from the various linear models. There were similar parental rankings across all models as evidenced by their correlations. Pearson

Table 3-4. Mean of estimated reliabilities ($\overline{\hat{r}_e^2}$) and true reliabilities ($\hat{r}_T^2$) from four mixed linear models for three mating designs and varying data structures.

Level of genetic control[a]	Model[b]	15-parent-half-diallel						52-parent-circular				105-parent Polymix			
		1[c]		2		3		1		2		1		2	
		$\overline{\hat{r}_e^2}$[d]	$\hat{r}_T^2$	$\overline{\hat{r}_e^2}$	$\hat{r}_T^2$	$\overline{\hat{r}_e^2}$	$\hat{r}_T^2$	$\overline{\hat{r}_e^2}$	$\hat{r}_T^2$	$\overline{\hat{r}_e^2}$	$\hat{r}_T^2$	$\overline{\hat{r}_e^2}$	$\hat{r}_T^2$	$\overline{\hat{r}_e^2}$	$\hat{r}_T^2$
1 (low)	I	0.687	0.710	0.616	0.632	0.491	0.466	0.562	0.556	0.485	0.467	0.537	0.549	0.418	0.419
$h^2 = 0.10$	III	0.821	0.725	0.732	0.648	0.579	0.488	0.666	0.565	0.568	0.479	0.629	0.545	0.493	0.428
$r_B = 0.60$	IV	0.750	0.716	0.702	0.641	0.628	0.494	0.743	0.567	0.666	0.485	---		---	
$\gamma = 1.00$	V	0.894	0.722	0.835	0.651	0.743	0.507	0.865	0.570	0.773	0.489	---		---	
2 (moderate)	I	0.792	0.811	0.744	0.765	0.627	0.641	0.724	0.727	0.661	0.655	0.703	0.709	0.598	0.605
$h^2 = 0.20$	III	0.873	0.814	0.819	0.771	0.688	0.648	0.789	0.730	0.717	0.657	0.762	0.710	0.641	0.608
$r_B = 0.75$	IV	0.829	0.812	0.799	0.768	0.738	0.654	0.841	0.729	0.784	0.658	---		---	
$\gamma = 0.50$	V	0.911	0.814	0.877	0.772	0.807	0.658	0.911	0.730	0.846	0.659	---		---	
3 (strong)	I	0.872	0.882	0.838	0.851	0.746	0.761	0.839	0.843	0.780	0.784	0.801	0.803	0.706	0.713
$h^2 = 0.30$	III	0.899	0.882	0.864	0.851	0.769	0.763	0.863	0.844	0.800	0.784	0.823	0.804	0.726	0.713
$r_B = 0.90$	IV	0.891	0.882	0.868	0.851	0.813	0.767	0.906	0.844	0.855	0.785	---		---	
$\gamma = 0.25$	V	0.919	0.882	0.895	0.852	0.839	0.768	0.930	0.844	0.878	0.785	---		---	

Notations for a, b and c are the same as those in table 3-2.

d. Each mean is calculated based on approximately 7,500, 26,000 and 50,000 estimates respectively for half-diallel, circular and polymix mating designs respectively.

e. Standard errors of the means varies between 0.000 and 0.002.

correlation coefficients among the predicted breeding values from the 5 analytical models for any data set were always close to 1.

Reliability of prediction

Estimates of reliability generally followed the same pattern as those of heritability estimates and predicted genetic gains. Reliability estimates from full analytical models were closest to the values of true reliability, especially for unbalanced data sets in the genetic architecture level I, II, and III. In contrast, most reliability estimates from incomplete models were significantly biased upward (Table 3-4), which indicated that the predicted BVs were stated to be more reliable than they truly are. The differences between true and estimated reliability were largest for Model V with severely unbalanced data in the genetic architecture level 1 and these biases became substantially smaller for balanced data in the genetic architecture level 3. Values of true reliability were only significantly affected by the true population genetic parameters and the degrees of data imbalance, but differed little by the full and incomplete analytical models for a given data set. Estimates of reliability, on the other hand, were affected greatly not only by true population genetic architectures and the degrees of data imbalance, but also by the analytical models.

Goodness of Approximation of Biases for Unbalanced Data

Results from analyses of simulated data using REML approach for variance component estimation and then BLUP for breeding value prediction have confirmed that, for balanced data created from the polymix and half-diallel mating designs, the magnitudes of empirical biases were identical to those calculated from the theoretical formulae (Table 3-5,

3-6, and 3-7). For unbalanced data, the empirical biases were generally greater than that calculated from the formulae using the average design parameters of experiments. This trend was especially strong for the incomplete linear models ignoring SCA effect and its interaction with environment, which suggested that the real biases from incomplete linear model for unbalanced data were generally inadequately appreciated using the theoretical approximations with formulae derived based on balanced data (Table 3-5 and 3-6). However, moderate to high Pearson correlation coefficients existed between the empirical and theoretically approximated biases (Table 3-7). Simple regressions of approximated biases on realized empirical biases were significant ($p<0.0001$) for all data structures considered in this study. The approximation methods worked much better for the incomplete model ignoring the effects of G x E interactions than for the incomplete model ignoring SCA effects. For instance, for incomplete models ignoring G x E interactions, the theoretically approximated biases have accounted for more than 80% of the variation of the empirical biases over repeated random samples of a given genetic architecture, mating design and data imbalance. In contrast, for incomplete models ignoring SCA effects, the theoretically approximated biases only accounted for 41-78% of the variation in the empirical biases, depending on the properties of unbalanced data.

For the seemingly balanced full-sib data created from circular mating designs, the approximated biases only matched the empirical biases perfectly for incomplete models ignoring all G x E interactions. For all other incomplete models ignoring SCA or/and its interaction with environment, the theoretical biases did not perfectly match the empirical

Table 3-5. Mean empirical ($\bar{b}_{VE}$) and theoretical ($\bar{b}_{VT}$) biases of the estimates of additive genetic variances from incomplete linear models expressed as the percentage to the estimates of additive genetic variances from the full models (i.e., $\bar{b}_V=\Delta\sigma^2_{gca}/\sigma^2_{gca}*100\%$)

Level of additive genetic control[a]	Model[b]	15-parent-half-diallel						52-parent-circular				105-parent Polymix			
		Balanced		Unbalanced 1		Unbalanced 2		Balanced		Unbalanced		Balanced		Unbalanced	
		$\bar{b}_{VE}$	$\bar{b}_{VT}$	$\bar{b}_{VE}$	$\bar{b}_{VT}$	$\bar{b}_{VE}$	$\bar{b}_{VT}$	$\bar{b}_{VE}$	$\bar{b}_{VT}$	$\bar{b}_{VE}$	$\bar{b}_{VT}$	$\bar{b}_{VE}$	$\bar{b}_{VT}$	$\bar{b}_{VE}$	$\bar{b}_{VT}$
1 (low)	II	-1.27	-1.27	-1.88	-1.85	-3.56	-3.44	-4.25	-4.13	-4.43	-4.52				
$h^2=0.10$	III	16.47	16.47	16.41	16.52	15.96	16.52	17.00	17.00	16.75	16.86	16.00	16.00	15.84	15.24
$r_B=0.60$	IV	7.62	7.62	12.34	11.16	29.55	23.33	31.63	25.45	39.16	29.64				
$\gamma=1.00$	V	24.87	24.87	30.33	28.83	48.21	41.15	52.05	45.88	59.78	50.63				
2 (moderate)	II	-0.30	-0.30	-0.48	-0.47	-1.13	-1.09	-1.19	-1.15	-1.55	-1.55				
$h^2=0.20$	III	8.38	8.38	8.28	8.27	8.29	8.33	8.38	8.38	8.03	8.03	8.04	8.04	8.00	7.62
$r_B=0.75$	IV	3.86	3.86	6.33	5.70	16.08	11.36	15.80	12.77	18.59	14.12				
$\gamma=0.50$	V	12.43	12.43	14.98	14.19	24.82	20.06	24.86	21.86	27.54	22.99				
3 (strong)	II	-0.05	-0.05	-0.08	-0.09	-0.15	-0.20	-0.16	-0.16	-0.29	-0.39				
$h^2=0.30$	III	2.76	2.76	2.71	2.70	2.63	2.63	2.72	2.72	2.47	2.47	2.66	2.66	2.67	2.54
$r_B=0.90$	IV	1.88	1.88	3.13	2.75	7.89	5.78	7.64	6.26	9.84	7.06				
$\gamma=0.25$	V	4.68	4.68	5.95	5.49	10.95	8.61	10.37	9.01	12.61	9.93				

a. $h^2=4\sigma_{gca}{}^2/\sigma_p{}^2$; $r_B=\sigma_{gca}{}^2/(\sigma_{gca}{}^2+\sigma_{ge}{}^2)$; $\gamma=\sigma_{sca}{}^2/\sigma_{gca}{}^2$.

b. Types of analytical linear models: II) SCA x E interaction excluded from full model, III) SCA x E and GCA x E interactions excluded, IV) SCA and SCA x E interactions excluded, and V) only additive genetic effect considered.

c. Types of data imbalance: for half-diallel design, unbalanced 1=35 missing crosses and 40% mortality, unbalanced 2= 70 missing crosses and 40% mortality; for circular design, unbalanced =13 missing crosses and 40% mortality; for polymix design, unbalanced = 40% mortality.

d 500 random experiments were simulated.

Table 3-6 . Mean empirical ($\bar{b}_{HE}$) and theoretically approximated ($\bar{b}_{HT}$) biases of estimates of heritability from incomplete linear models expressed as the percentage to the estimates of heritability from the full models (i.e., $\bar{b}_h = \Delta h^2/h^2 * 100\%$)

Level of additive genetic control	Model	15-parent-half-diallel						52-parent-circular				105-parent Polymix			
		Balanced		Unbalanced 1		Unbalanced 2		Balanced		Unbalanced		Balanced		Unbalanced	
		$\bar{b}_{HE}$	$\bar{b}_{HT}$	$\bar{b}_{HE}$	$\bar{b}_{HT}$	$\bar{b}_{HE}$	$\bar{b}_{HT}$	$\bar{b}_{HE}$	$\bar{b}_{HT}$	$\bar{b}_{HE}$	$\bar{b}_{HT}$	$\bar{b}_{HE}$	$\bar{b}_{HT}$	$\bar{b}_{HE}$	$\bar{b}_{HT}$
1 (weak)	II	-1.31	-1.31	-1.86	-1.86	-3.72	-3.56	-4.38	-4.16	-4.52	-4.60				
$h^2 = 0.10$	III	15.08	15.08	15.04	15.14	14.54	15.02	15.26	15.26	15.12	15.24	16.00	16.00	15.84	15.84
$r_B = 0.60$	IV	7.65	7.65	12.26	11.45	28.70	22.11	31.16	25.46	38.38	29.54				
$\gamma = 1.00$	V	25.67	25.67	31.12	29.65	48.50	41.55	51.77	46.10	59.32	50.17				
2 (moderate)	II	-0.30	-0.30	-0.57	-0.52	-1.21	-1.16	-1.21	-1.17	-1.61	-1.57				
$h^2 = 0.20$	III	7.19	7.19	7.16	7.11	6.74	6.74	6.80	6.80	6.50	6.49	8.04	8.04	8.00	7.62
$r_B = 0.75$	IV	3.87	3.87	6.29	5.70	16.10	11.58	15.44	12.78	18.07	14.26				
$\gamma = 0.50$	V	12.95	12.95	15.51	14.72	25.48	20.77	24.67	22.03	27.14	23.14				
3 (strong)	II	-0.05	-0.05	-0.07	-0.10	-0.17	-0.21	-0.16	-0.17	-0.29	-0.39				
$h^2 = 0.30$	III	2.12	2.12	2.08	2.04	2.07	2.04	1.94	1.94	1.76	1.77	2.66	2.66	2.67	2.54
$r_B = 0.90$	IV	1.89	1.89	3.09	2.75	8.22	5.78	7.41	6.27	9.44	7.07				
$\gamma = 0.25$	V	4.65	4.65	6.13	5.69	11.33	8.88	10.03	9.87	12.27	9.90				

Notations are the same as in Table 3-5.

Table 3-7 . Regression slopes (b) and r-squares (R^2) of approximated biases on the empirical biases for the estimates of additive genetic variances from different incomplete linear models.

Level of additive genetic control[a]	Model	15-parent-half-diallel				52-parent-circular			
		Unbalanced 1		Unbalanced 2		Balanced		Unbalanced	
		b[¶]	R^2	b	R^2	b	R^2	b	R^2
1 (low)	II	0.726	0.795	0.816	0.855	0.903	0.949	1.061	0.955
$h^2 = 0.10$	III	0.897	0.885	0.845	0.797	1.000	1.000	0.922	0.907
$r_B = 0.60$	IV	0.508	0.549	0.484	0.661	0.610	0.770	0.612	0.815
$\gamma = 1.00$	V	0.699	0.712	0.532	0.654	0.645	0.782	0.629	0.804
2 (moderate)	II	0.651	0.746	0.743	0.858	0.940	0.965	0.914	0.908
$h^2 = 0.20$	III	0.864	0.880	0.836	0.821	1.000	1.000	0.837	0.841
$r_B = 0.75$	IV	0.485	0.536	0.402	0.411	0.611	0.757	0.618	0.785
$\gamma = 0.50$	V	0.676	0.723	0.469	0.594	0.648	0.771	0.638	0.787
3 (strong)	II	0.780	0.829	0.801	0.825	1.010	0.975	0.971	0.932
$h^2 = 0.30$	III	0.838	0.847	0.823	0.803	1.000	1.000	0.900	0.897
$r_B = 0.90$	IV	0.467	0.540	0.485	0.697	0.648	0.748	0.555	0.752
$\gamma = 0.25$	V	0.588	0.653	0.478	0.669	0.673	0.774	0.576	0.762

[¶] 500 randomly simulated data samples were used in each of the regression analysis.

biases although they are highly correlated (Table 3-7). This indicated that data created from the circular mating designs are not completely balanced even if there were no missing values according to the mating design.

Discussion

Bias

Simulations in this study with unbalanced data of different mating designs, genetic architecture have revealed considerable biases in heritability estimates and genetic gain prediction from the use of incomplete analytical mixed linear models when dominance effect and G x E interaction were present in the experimental data. Results from this study are consistent with those from practical data analyses. For instance, in forestry, Dieters et al. (1995) found that the average of heritability estimates from single test (ignoring G x E

interaction) were significantly higher than those from multiple-test analyses for a large number of full-sib slash pine (*Pinus elliottii* Engelm var. *elliottii*) families even when dominance effects were considered in the analytical models. In poultry breeding, Wei et al. (1993) found that an additive model ignoring dominance effect significantly increased the estimates of heritabilities as compared with an animal model that included dominance effects. Uimari and Mäntysaari (1993) indicated that estimated reliabilities from models not considering dominance effects were substantially higher than the empirical reliability estimates based on the correlation between pedigree index and final sire proof in Finnish dairy cow evaluation. Quinton and Smith (1997) further indicated that heritability estimated from an additive model was too high to predict breeding values and genetic change when an empirical check was conducted with a large body of Canadian pig performance records to check on the predicted benefits of BLUP in genetic evaluation.

The biases in heritability estimates and predicted genetic gains are the consequences of the over-estimation of additive genetic variances from the incomplete analytical models. For a given level of the genetic architectures and data imbalance in this study, it was observed that the mean of estimated additive genetic variances from 500 independent experiments always converged to a larger value than the true additive genetic variance for incomplete models but not for the full analytical models. The average of total phenotypic variances, however, converged close to the designed true value 10 for all models. Based on the theoretically well-established relationships among additive genetic variance, heritability and genetic gain (Falconer 1981), it is thus not surprising to observe such biases in heritability estimates and predicted genetic gains from incomplete models.

Biases in heritability estimates and predicted genetic gains by Model V (no G x E and SCA) were approximately the sum of the biases produced by Model III (no G x E) and IV (no SCA). This trend was observed in the two full-sib mating designs across all levels of genetic architecture and data imbalance. This implies that biases from incomplete models behave in a cumulative manner, that is, the more effects that are ignored from an analytical model, the larger the biases in heritability estimates and predicted genetic gains are. This result is also consistent with the theoretical considerations based on balanced data.

The relatively stable biases from Model III across mating designs and data imbalance is attributable to the constant field experimental design (*i.e.*, 4 locations, 15 blocks/location) in this study. This is completely in agreement with the theoretical formula of bias for this incomplete model, which is mainly a function of the number of locations in genetic tests. Thus, there should be no evidence to suggest that the magnitude of biases from Model III will be limited to the scale involved in this study. Therefore, the biases from Model III may change dramatically given different field experimental designs, such as the numbers of locations.

Approximation

The theoretical formulae of biases derived based on balanced data in chapter 2 did not estimate the magnitudes of biases exactly as they were for unbalanced data. Actual biases for unbalanced data were mostly larger than those estimated from the formulae for incomplete mixed model III, IV, and V. This trend was especially strong for severely unbalanced data with weak additive genetic control over a trait (Table 3-5 and 3-6). The larger actual biases from analysis of unbalanced data indicated that those approximation

formulae of biases derived based on balanced data were inadequate. This discrepancy may be caused by the worsened orthogonality among experimental factors in the unbalanced experimental data, which resulted in the different ways of pooling sum of squares in the incomplete models and different expected mean squares in data analysis.

Despite the inadequacy of the approximation formulae for severely unbalanced data, there were high correlations between the approximated biases for each of the incomplete mixed linear models and their actual biases. This, on the other hand, suggests that the approximated biases using the average design parameters of an experiments can be viewed as the minimum bias that may be expected from using incomplete linear models in forest genetic data analysis.

Implications for tree-breeding programs

Many economically important traits of forest trees have low heritabilities and appreciable dominance effects and G x E interaction. For instance, in several conifer species, narrow-sense heritability for volume growth has been reported to range from 0.1 to 0.2 (Yeh and Heaman 1987; Dieters et al. 1995; Li et al. 1996). In loblolly pine (*Pinus taeda* L.), dominance variance has been estimated to account for 50% - 70% of the total genetic variance in height, 50%-55% in DBH and 45% - 60% in volume (Li et al. 1996). In slash pine, the ratio of dominance variance to additive variance was estimated to range from 0.32 to 0.60 with multi-site data and 0.54 - 0.68 with single site data, and type B genetic correlations among sites were estimated to range from 0.61 to 0.76 except for a few cases in which estimates were as high as 0.88 (Dieters et al. 1995). In longleaf pine, type-B genetic correlations were also estimated to be in the range of 0.61 to 0.74 (Adams et al. 1994). These

well-estimated genetic parameters from large data samples have indicated that genetic architectures in most conifer tree breeding programs would fall into the categories of genetic architecture level 1 and 2 as simulated in this study. Furthermore, a 15-parent-half-diallel design is impractical. More commonly, the number of crosses per parent would be close to the circular mating designs, in which 4-6 crosses per parent can be made. Combining the potential genetic architectures and data structure in forest genetic tests suggests that estimates of heritability, reliability and predictions of genetic gains from incomplete models would be subject to severe biases, especially for those models that ignore both dominance effects and G x E interaction.

Because true genetic gain from selection would not suffer from using incomplete analytical model as indicated in this study, incomplete analytical model considering only additive genetic effects could be used to save computational costs in cases where an accurate prediction of genetic gain is not critical. However, whenever accurate estimation of heritability and prediction of genetic gain are required, incomplete models should be avoided, especially incomplete models that ignore both dominance genetic effects and G x E interaction.

Conclusion

While full analytical mixed models consistently yield unbiased estimates of heritability and accurate predictions of genetic gain, incomplete mixed models generally yield serious biases when complex genetic structures are present in the data. The magnitudes of biases from incomplete models are considerably larger when the population true

heritability is low, G x E is large and only a few crosses per parent are available. Reliability estimates from incomplete models generally give false indications of the accuracy of the breeding value prediction unless a trait is highly additively genetically controlled. Incomplete models, however, can be used to rank parents for selection as accurately as full-models. No evidence here suggests that true genetic gains suffer from using incomplete analytical models.

Approximation formulae of biases based on balanced data generally under-estimate the biases for incomplete mixed linear models when data are severely unbalanced. However, approximated biases using the average designed parameters of an unbalanced experiment can be safely used as an indication of the minimum bias caused by an incomplete mixed linear model, which can be used to evaluate the suitability of an incomplete linear model for a specific data structure.

CHAPTER 4
ESTIMATING TYPE B GENETIC CORRELATIONS WITH UNBALANCED DATA AND HETEROGENEOUS VARIANCES

Introduction

Defined as the genetic correlation of the same trait measured in different environments (Dickerson 1962; Yamada 1962; Burdon 1977), type B genetic correlations have many applications in tree improvement programs. In numerous genetic studies, estimates of type B genetic correlations have been used as quantitative measures of genotype-by-environment (G x E) interactions (Burdon 1977; Johnson and Burdon 1990; Woolaston et al. 1991; Adams et al. 1994; Cooper and Delacy 1994; Dieters et al. 1995; Dieters 1996; Pswarayi et al. 1997), which are important considerations in formulating breeding strategies and in deploying genetically improved materials (Zobel and Talbert 1984; White et al. 1993). In other tree improvement applications, estimates of type B genetic correlations frequently serve as the link for predicting genetic responses from indirect selection (Jiang 1985; White and Hodge 1989; Surles 1993; Wu 1993; Johnson 1997).

Statistical methods for estimating type B genetic correlation have been well established for balanced data. In forest genetic data analyses, type B genetic correlations have been routinely estimated using either the method of Yamada (1962) or the formula of Burdon (1977). So long as data are balanced, both methods are theoretically well defined and

yield identical and unbiased results. However, when severely unbalanced data are involved along with heterogeneous (genetic or/and environmental) variances, theoretical concerns arise about the general utilities of these approaches (Fernando et al. 1984) due to their potential biases.

The objectives of this study are twofold: (i) to develop a new approach for estimating type B genetic correlations that more properly accounts for unbalanced data with heterogeneous variances as well as different experimental designs across environments, and (ii) to compare numerically the estimates of type B genetic correlations from different methods using simulated data sets which have known population parameters.

Theoretical Considerations of Type B Genetic Correlations

Background

The concept of genetic correlation for the same trait measured in different environment was first used by Falconer (1952) and the theory was further developed by others (Robertson 1959; Dickerson 1962; Yamada 1962). To distinguish this type of genetic correlation from the genetic correlation between different traits measured on the same individuals, Burdon (1977) called the former type B genetic correlation.

From its definition, Yamada (1962) derived an estimation method of type B genetic correlation based on a two-way analysis of variance (ANOVA). For a pair of environments, the type B genetic correlation is estimated as:

$$\hat{r}_B = \frac{\hat{\sigma}_g^2}{\hat{\sigma}_g^2 + \hat{\sigma}_I^2 - \frac{(\hat{\sigma}_{g1} - \hat{\sigma}_{g2})^2}{2}} \qquad 4\text{-}1$$

where $\hat{r}_B$ is the estimate of type B genetic correlation, $\hat{\sigma}^2_g$ is the genetic variance component estimated from a two-way analysis of variance involving data from two environments assuming homogeneous variances between them, $\hat{\sigma}^2_I$ is the estimate of variance component for the effect of G x E interaction, $\hat{\sigma}^2_{g1}$ and $\hat{\sigma}^2_{g2}$ are the estimates of genetic variance components within environment 1 and 2, respectively.

Fernando et al. (1984) showed that for balanced data, Yamada's method is well defined. However, for unbalanced data Eq.4-1 yields biased estimates of type B genetic correlation under customary linear models (i.e., with the assumption of zero-covariance between random effects) unless genetic and environmental variances are identical across environments. While trying to theoretically justify Eq.1 with an alternative model, Itoh and Yamada (1990) acknowledged that, in the reparameterization of Eq.4-1, giving equal weights to each of the two environments is unreasonable if population sizes are finite and unequal. These questions and the possible violation of the assumption of homogeneous variances in a two-way analysis of variance have caused theoretical concerns about the biases that may result from the use of Yamada's method when data are severely unbalanced and variances are heterogeneous. However, the severity of such biases has not been well demonstrated based on empirical evidence of genetic testing data.

Burdon (1977) provided an alternative formula for the estimation of type B genetic correlation as:

$$\hat{r}_B = \frac{\hat{r}_{xy}}{\hat{h}_x \hat{h}_y} \qquad 4\text{-}2$$

where $\hat{r}_{xy}$ is the phenotypic correlation between genetic group means in environments **x** and **y**, $\hat{h}_{\bar{y}}$ and $\hat{h}_{\bar{y}}$ are square-roots of heritabilities of the genetic group means in environments **x** and **y,** respectively.

Burdon's formula is based on the assumption of uncorrelated non-genetic effects between two environments. Provided that no common environmental effects exist between two testing environments, this assumption would be theoretically appropriate. It was shown (Burdon 1977) that Eq.4-2 yields identical estimates of type B genetic correlations to that from Eq.4-1 when data are balanced. Since Burdon's method does not require a two-way analysis of variance, it tactically avoids the violation of the homogeneous variance assumption in cases where variances (either genetic or environmental) are heterogeneous across environments. In addition, because only genetic group means are used in the calculation of phenotypic correlation, and heritabilities of genetic group means are estimated separately in each environment, Eq.4-2 facilitates the calculation of type B genetic correlations in cases where experimental designs differ across genetic tests.

Despite of these obvious advantages, estimates of type B genetic correlations from Eq.4-2 may also become biased when missing values occur due to mortality or other unforeseeable factors in the experiments which cause unequal number of observations per genetic group. In such cases, genetic group means may be confounded by other experimental factors, resulting in biased estimates of genetic covariance. Again, the severity of the potential bias has not been numerically evaluated.

A New Approach Using BLUP-predicted Parental GCA Effects (GCA Approach)

Theory of the new approach

To simplify the derivation without losing generality, let $\bar{x}_{1.}, \bar{x}_{2.}, ..., \bar{x}_{i.}, ..., \bar{x}_{m.}$ be the means of half-sib progenies from *m* independent female parents measured in environment **x** from a randomized complete block (RCB) experimental design with one individual per family in each plot (single-tree plots). Similarly, let $\bar{y}_{i.}$ be the means of half-sib progenies of the same *m* parents measured in environment **y** with the same experimental design. Let n_{xi} and n_{yi} be, respectively, the number of offspring on which $\bar{x}_{i.}$ and $\bar{y}_{i.}$ are based.

For a one-way analysis of variance in each single environment, the analytical linear models are:

$$x_{ij}=\mu_x+\beta_{xj}+g_{xi}+e_{xij} \qquad 4\text{-}3a$$

and

$$y_{ik}=\mu_y+\beta_{yk}+g_{yi}+e_{yik} \qquad 4\text{-}3b$$

respectively for environment **x** and **y**;

where μ_x and μ_y are the overall means;

β_{xj} and β_{yk} are the fixed effects of block, $j=1, 2, ..., n_j$, and $k=1, 2, ..., n_k$;

g_{xi} and g_{yi} are the random effects of the ith family, $g_{xi} \sim \text{NID}(0, \sigma^2_{gx})$, $g_{yi} \sim \text{NID}(0, \sigma^2_{gy})$, $i= 1, 2, ..., m$;

and e_{xij} and e_{yik} are the jth (or kth) residual effect within the ith family, $e_{xij} \sim \text{NID}(0, \sigma^2_{xe})$, $e_{yik} \sim \text{NID}(0, \sigma^2_{ye})$;

Thus, $Var(\bar{x}_{i.})=var(\mu_x+\bar{\beta}_{.}+g_{xi}+\bar{e}_{xi.}) = \sigma^2_{gx}+\frac{\sigma^2_{ex}}{n_{xi}}$, and similarly, $Var(\bar{y}_{i.})=\sigma^2_{gy}+\frac{\sigma^2_{ey}}{n_{yi}}$.

Note when n_{xi} differs among families, $Var(\bar{x}_{i.})$ is heterogeneous among $\bar{x}_{i.}$'s, and this is also true among $Var(\bar{y}_{i.})$'s if $n_{yi} \neq n_y$=constant. Therefore, the phenotypic correlation between genetic group means from two environments, *i.e.*, $\hat{r}_{xy}=\frac{Cov(\bar{x}_{i.}\bar{y}_{i.})}{\sqrt{Var(\bar{x}_{i.})Var(\bar{y}_{i.})}}$, is ambiguously defined due to the different distributional properties among the $\bar{x}_{i.}$'s in environment **x** or/and among the $\bar{y}_{i.}$'s in environment **y**.

A reasonable approach to alleviate these heterogeneous variances among the $\bar{x}_{i.}$'s and $\bar{y}_{i.}$'s is to conduct a data standardization. Let

$$\bar{x}'_{i.}=\frac{(\bar{x}_{i.}-\mu_x-\bar{\beta}_{x.})}{\sqrt{Var(\bar{x}_{i.})}}=\frac{(\bar{x}_{i.}-\mu_x-\bar{\beta}_{x.})}{\sqrt{\sigma^2_{gx}+\frac{\sigma^2_{ex}}{n_{xi}}}}, \qquad \text{and } \bar{y}'_{i.}=\frac{(\bar{y}_{i.}-\mu_y-\bar{\beta}_{y.})}{\sqrt{Var(\bar{y}_{i.})}}=\frac{(\bar{y}_{i.}-\mu_y-\bar{\beta}_{y.})}{\sqrt{\sigma^2_{gy}+\frac{\sigma^2_{ey}}{n_{yi}}}} \qquad 4\text{-}4$$

So that $Var(\bar{x}'_{i.})=Var(\bar{y}'_{i.})=1$.

By the central limit theorem, when n_{xi} and n_{yi} are sufficiently large, then

$$\bar{x}_{i.} \sim N(\mu_x+\bar{\beta}_{x.}, \sigma^2_{gx}+\frac{\sigma^2_{ex}}{n_{xi}}), \qquad \bar{x}'_{i.} \sim N(0, 1);$$

and

$$\bar{y}_{i.} \sim N(\mu_y+\bar{\beta}_{y.}, \sigma^2_{gy}+\frac{\sigma^2_{ey}}{n_{yi}}), \qquad \bar{y}'_{i.} \sim N(0, 1).$$

Thus, $\hat{r}=\frac{Cov(\bar{x}'_{i.}\bar{y}'_{i.})}{\sqrt{Var(\bar{x}'_{i.})Var(\bar{y}'_{i.})}}=Cov(\bar{x}'_{i.}\bar{y}'_{i.})$ is uniformly defined.

By conducting such data standardization, two major objectives are achieved, i.e., fixed effects of blocks are removed and variances are homogenized among families within an environment.

Let $n_x = \sum_{i=1}^{m} \frac{n_{xi}}{m}$ and $n_y = \sum_{i=1}^{m} \frac{n_{yi}}{m}$, and note that $Var(\bar{x}_{i.}')=Var(\bar{y}_{i.}')=1$, thus

$$\hat{r}=Cov(\bar{x}_{i.}',\bar{y}_{i.}')=Cov(\frac{\bar{x}_{i.}}{\sqrt{\sigma_{gx}^2+\frac{\sigma_{ex}^2}{n_{xi}}}},\frac{\bar{y}_{i.}}{\sqrt{\sigma_{gy}^2+\frac{\sigma_{ey}^2}{n_{yi}}}})\frac{\sqrt{\sigma_{gx}^2+\frac{\sigma_{ex}^2}{n_x}}}{\sqrt{\sigma_{gx}^2+\frac{\sigma_{ex}^2}{n_x}}}\frac{\sqrt{\sigma_{gy}^2+\frac{\sigma_{ey}^2}{n_y}}}{\sqrt{\sigma_{gy}^2+\frac{\sigma_{ey}^2}{n_y}}}$$

$$=Cov(\frac{\bar{x}_{i.}\sqrt{\sigma_{gx}^2+\frac{\sigma_{ex}^2}{n_x}}}{\sqrt{\sigma_{gx}^2+\frac{\sigma_{ex}^2}{n_{xi}}}},\frac{\bar{y}_{i.}\sqrt{\sigma_{gy}^2+\frac{\sigma_{ey}^2}{n_y}}}{\sqrt{\sigma_{gy}^2+\frac{\sigma_{ey}^2}{n_{yi}}}})/(\sqrt{\sigma_{gx}^2+\frac{\sigma_{ex}^2}{n_x}}\sqrt{\sigma_{gy}^2+\frac{\sigma_{ey}^2}{n_y}})$$

Let $w_{xi}=\frac{\sqrt{\sigma_{gx}^2+\frac{\sigma_{ex}^2}{n_x}}}{\sqrt{\sigma_{gx}^2+\frac{\sigma_{ex}^2}{n_{xi}}}}$, and $w_{yi}=\frac{\sqrt{\sigma_{gy}^2+\frac{\sigma_{ey}^2}{n_y}}}{\sqrt{\sigma_{gy}^2+\frac{\sigma_{ey}^2}{n_{yi}}}}$, then

$$\hat{r}=Cov(\bar{x}_{i.}w_{xi},\bar{y}_{i.}w_{yi})/(\sqrt{\sigma_{gx}^2+\frac{\sigma_{ex}^2}{n_x}}\sqrt{\sigma_{gy}^2+\frac{\sigma_{ey}^2}{n_y}})$$

$$\approx \hat{\sigma}_{gxy}/(\sqrt{\sigma_{gx}^2+\frac{\sigma_{ex}^2}{n_x}}\sqrt{\sigma_{gy}^2+\frac{\sigma_{ey}^2}{n_y}}), \qquad 4\text{-}5$$

This results because residual effects across environments are assumed to be uncorrelated as in Eq.4-2, and when data are balanced, $w_{xi}=w_{yi}=1$, so that $Cov(\bar{x}_{i.},\bar{y}_{i.})$ estimates one-quarter of the additive genetic covariance. For unbalanced data, w_{xi} and w_{yi} can be viewed as the weights used in the calculation of genetic covariance to account for the unequal number of observations among genetic groups. In the above calculation, the means of genetic groups based on fewer observations will have less leverage and, therefore,

contribute less to the calculation of genetic covariance than the means of genetic groups based on more observations. The significance of Eq.4-5 is when data are balanced it yields identical results to those from Eq.4-1 and Eq.4-2, but when data are unbalanced it yields properly weighted phenotypic correlation between family means in two environments after removing fixed effects.

From Eq.4-5, it can be further developed that

$$\hat{r}_B = \hat{r}\sqrt{\sigma^2_{gx} + \frac{\sigma^2_{ex}}{n_x}}\sqrt{\sigma^2_{gy} + \frac{\sigma^2_{ey}}{n_y}} \Big/ (\hat{\sigma}_{gx}\,\hat{\sigma}_{gy}) = \frac{\hat{r}}{\hat{h}_{\bar{x}}\hat{h}_{\bar{y}}}, \qquad 4\text{-}6$$

where $\hat{r}_B$ is the estimate of type B genetic correlation, $\hat{h}_{\bar{x}}$ and $\hat{h}_{\bar{y}}$ are square-root heritabilities of the genetic group means in environments **x** and **y,** respectively.

Equation 4-6 appears identical to the format of Burdon in Eq.4-2 except that $\hat{r}$ is estimated differently from $\hat{r}_{xy}$ in Eq.4-2. In the current approach, $\hat{r}$ is estimated based on residuals after removing all fixed effects and has been adjusted for the unequal number of observations of genetic group means during the procedure of data standardization.

Operational calculations

Equations 4-5 and 4-6 are theoretically informative, but computationally tedious, especially when complex experimental designs are involved. We show below that the required calculation is best accomplished by using the predicted parental GCA effects with the technique of best linear unbiased prediction (BLUP).

Let $\hat{g}_{x1}$, $\hat{g}_{x2}$, ...,$\hat{g}_{xm}$ be the predicted parental GCA effects of the *m* parents in environment **x** and $\hat{g}_{y1}$, $\hat{g}_{y2}$, ...,$\hat{g}_{ym}$ be the predicted GCA effects of the same *m* parents in

environment **y**. From the theory of BLUP (Henderson 1984), it is known that $\hat{g}_{xi}=b_{\bar{x}i}(\bar{x}_i-\hat{\tau}_{xi})$ and $\hat{g}_{yi}=b_{\bar{y}i}(\bar{y}_i-\hat{\tau}_{yi})$, where $b_{\bar{x}i}$ and $b_{\bar{y}i}$ are regression coefficients, and $\hat{\tau}_{xi}$ and $\hat{\tau}_{yi}$ are generalized least square estimates of the fixed effects corresponding to $\bar{x}_i$ and $\bar{y}_i$, respectively. With the above RCB experimental designs and considering block being fixed effects, it can be derived (Searle et al. 1992, p.68) that:

$$b_{\bar{x}i}=\frac{\sigma^2_{gx}}{\sigma^2_{gx}+\frac{\sigma^2_{ex}}{n_{xi}}} \quad \text{and} \quad b_{\bar{y}i}=\frac{\sigma^2_{gy}}{\sigma^2_{gy}+\frac{\sigma^2_{ey}}{n_{yi}}}. \qquad 4\text{-}7$$

Therefore, with the assumption of BLUP (Henderson 1984; White and Hodge 1989) that variance components are known without error, we have

$$Var(\frac{\hat{g}_{xi}}{\sqrt{b_{\bar{x}i}}})=\frac{1}{b_{\bar{x}i}}Var(b_{\bar{x}i}\bar{x}_i)=b_{\bar{x}i}Var(\bar{x}_i)$$

$$=(\frac{\sigma^2_{gx}}{\sigma^2_{gx}+\frac{\sigma^2_{ex}}{n_{xi}}})(\sigma^2_{gx}+\frac{\sigma^2_{ex}}{n_{xi}}) \qquad \text{(by substituting } b_{\bar{x}i} \text{ of Eq.4- 7)}$$

$$=\sigma^2_{gx}, \qquad 4\text{-}8a$$

and similarly, $Var(\frac{\hat{g}_{yi}}{\sqrt{b_{\bar{y}i}}})=\sigma^2_{gy}$. $\qquad 4\text{-}8b$

Thus $$Cov(\frac{\hat{g}_{xi}}{\sqrt{b_{\bar{x}i}}},\frac{\hat{g}_{yi}}{\sqrt{b_{\bar{y}i}}})=Cov(\sqrt{b_{\bar{x}i}}*\bar{x}_i\sqrt{b_{\bar{y}i}}*\bar{y}_i)$$

$$=Cov(\sqrt{\frac{\sigma^2_{gx}}{\sigma^2_{gx}+\frac{\sigma^2_{ex}}{n_{xi}}}}*\bar{x}_i,\sqrt{\frac{\sigma^2_{gy}}{\sigma^2_{gy}+\frac{\sigma^2_{ey}}{n_{yi}}}}*\bar{y}_i)$$

$$= \sigma_{gx}\sigma_{gy} Cov\left(\frac{\bar{x}_i}{\sqrt{\sigma^2_{gx} + \frac{\sigma^2_{ex}}{n_{xi}}}}, \frac{\bar{y}_i}{\sqrt{\sigma^2_{gy} + \frac{\sigma^2_{ey}}{n_{yi}}}}\right). \qquad 4\text{-}9$$

Let $$r^* = \frac{Cov\left(\frac{\hat{g}_{xi}}{\sqrt{b_{xi}}}, \frac{\hat{g}_{yi}}{\sqrt{b_{yi}}}\right)}{\sqrt{Var\left(\frac{\hat{g}_{xi}}{\sqrt{b_{xi}}}\right) Var\left(\frac{\hat{g}_{yi}}{\sqrt{b_{yi}}}\right)}} = \frac{\sigma_{gx}\sigma_{gy} Cov\left(\frac{\bar{x}_i}{\sqrt{\sigma^2_{gx} + \frac{\sigma^2_{ex}}{n_{xi}}}}, \frac{\bar{y}_i}{\sqrt{\sigma^2_{gy} + \frac{\sigma^2_{ey}}{n_{yi}}}}\right)}{\sigma_{gx}\sigma_{gy}}$$

$$= Cov\left(\frac{\bar{x}_i}{\sqrt{\sigma^2_{gx} + \frac{\sigma^2_{ex}}{n_{xi}}}}, \frac{\bar{y}_i}{\sqrt{\sigma^2_{gy} + \frac{\sigma^2_{ey}}{n_{yi}}}}\right). \qquad 4\text{-}10$$

Thus $r^* = \hat{r}$ (comparing Eq.4-5 with Eq.4-11). 4-11

In the theory of breeding value prediction, Mrode (1996) demonstrates that, in general, $b = kr_a^2$, where b is regression coefficient, k is an integer (for parental GCA effect, k=1) and r_a is the 'accuracy' of prediction with the technique of BLP (best linear prediction). In Appendix 4-1, it is shown that: $E[Cov(\frac{\hat{g}_{xi}}{r_{axi}}, \frac{\hat{g}_{yi}}{r_{ayi}})\overline{r_{ax} r_{ay}}] = \sigma_{gxy}$, $E[Var(\frac{\hat{g}_{xi}}{r_{axi}})] = \sigma^2_{gx}$, and $E[Var(\frac{\hat{g}_{yi}}{r_{ayi}})] = \sigma^2_{gy}$. Thus, the type B genetic correlation is estimated as:

$$\hat{r}_B = \frac{Cov\left(\frac{\hat{g}_{xi}}{\hat{r}_{axi}}, \frac{\hat{g}_{yi}}{\hat{r}_{ayi}}\right)}{\overline{\hat{r}_{ax}\hat{r}_{ay}}\sqrt{Var\left(\frac{\hat{g}_{xi}}{\hat{r}_{axi}}\right) Var\left(\frac{\hat{g}_{yi}}{\hat{r}_{ayi}}\right)}} = \frac{\hat{r}^*}{\overline{\hat{r}_{ax}\hat{r}_{ay}}}, \qquad 4\text{-}12$$

where $\hat{r}_B$ is the estimate of type B genetic correlation, r^* is Pearson correlation coefficient between weighted parental GCA effects predicted in environments **x** and **y** using BLUP, and $\overline{\hat{r}_{ax}\hat{r}_{ay}}$ is mean of products of prediction accuracies in the two environments.

It should be pointed out that in the derivation of Eq.4-12, $\hat{r}_{ai}$ was assumed equal to the prediction 'accuracy' using the technique of BLP. In practice, however, the technique of BLUP needs to be applied to remove fixed effects and prediction 'accuracy' is generally given smaller in BLUP than in BLP due to the fact that BLUP estimates fixed effects while BLP does not (White and Hodge 1989; Searle et al. 1992). The appropriate BLP $\hat{r}_{ai}$ thus has to be either precisely calculated using computer programs or converted approximately from the output of BLUP programs. To convert BLUP $\hat{r}_{ai}$ to BLP $\hat{r}_{ai}$, an adjustment factor $\sqrt{\frac{m}{m-1}}$ needs to be multiplied to the BLUP $\hat{r}_{ai}$, where m is the number of parents tested in the environment. Based on our numerical studies, the adjustment factor is satisfactory for a variety of balanced and moderately unbalanced data sets created from half-sib or full-sib mating designs. For severely unbalanced data, such as a circular mating design with more than 50 parents being used, it is best to calculate the BLP $\hat{r}_{ai}$ exactly using appropriate computer programs.

In summary, Eq.4-12 estimates type B genetic correlations using predicted parental GCA effects from two concerned environments and involves the following steps:

1. Predict parental GCA effects in each environment using the technique of BLUP (best linear unbiased prediction) with available computer BLUP software such as GAREML

(Huber 1993), MTDFREML (Boldman et al. 1995), ASREML (Gilmour et al. 1997), etc..

2. Divide each predicted parental GCA effect by its 'accuracy' of prediction in each environment to create weighted parental GCA effects, i.e., $\frac{\hat{g}_i}{\hat{r}_{ai}}$, where $\hat{g}_i$ is the predicted parental GCA effect of the ith parent and $\hat{r}_{ai}$ is its prediction 'accuracy' from BLP. If $\hat{r}_{ai}$ is the prediction 'accuracy' from BLUP, $\hat{r}_{ai}$ needs to be multiplied by an adjustment factor $\sqrt{\frac{m}{m-1}}$, where m is the number of parents tested in the environment.

3. Calculate the Pearson correlation coefficient of the adjusted parental GCA effects between the two concerned environments.

4. Divide the correlation coefficient from step (iii) by the mean of the products of the prediction accuracies for parents in the two environments.

The advantage of this approach lies in the fact that by using predicted parental GCA effects in calculating type B genetic correlation several goals are achieved simultaneously: (1) fixed effects are properly removed with the technique of best linear unbiased estimation (BLUE), so that they no longer confound genetic effects when data are unbalanced; (2) heterogeneous variances associated with genetic group means due to their unequal number of observations and different precision are automatically adjusted for in the process of breeding value prediction; (3) this method can be extended to calculate additive type B genetic correlations when data of full-sib progenies are used; and (4) this method is applicable when different experimental designs are used in the two environments, such as different plot sizes.

Numerical Comparisons of Estimation Methods

Methods

Data generation

Simulated data were used in the numerical comparison because the true underlining type B genetic correlation was known for each data set. Data were generated based on a randomized complete block design with single-tree plots, which is recommended by several studies in forest genetic testing (Lambeth and Gladstone 1983; Loo-Dinkins and Tauer 1987; Loo-Dinkins et al. 1990; White 1996). Genetic structures of the data were based on half-sib families created from a polymix mating design with 100 female parents. The linear model in matrix notation is:

$$y=\begin{bmatrix} y_1 \\ y_2 \end{bmatrix}=\begin{bmatrix} 1_1\mu_1 \\ 1_2\mu_2 \end{bmatrix}+\begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix}\begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}+\begin{bmatrix} Z_1 & 0 \\ 0 & Z_2 \end{bmatrix}\begin{bmatrix} g_1 \\ g_2 \end{bmatrix}+\begin{bmatrix} Z_{m1} & 0 \\ 0 & Z_{m2} \end{bmatrix}\begin{bmatrix} g_{m1} \\ g_{m2} \end{bmatrix}+\begin{bmatrix} e_1 \\ e_2 \end{bmatrix} \qquad 4\text{-}13$$

where y_i is the n_i *x 1* vector of phenotypic observations in environment *i, i=1,2*; n_i is the number of observations in environment *i*; μ_i is the overall mean in environment *i* and 1_i is an n_i *x 1* vector of 1s; $\mathbf{X_i}$ is the incidence matrix relating the block effects (vector $\boldsymbol{\beta}_i$) in environment *i;* $\mathbf{Z_i}$ is the incidence matrix relating to the female genetic effects (vector $\boldsymbol{g_i}$) in environment *i;* $\mathbf{Z_{mi}}$ is the incidence matrix relating to the genetic effects of Mendelian sampling (vector $\boldsymbol{g_{mi}}$) in environment *i*; $\boldsymbol{e_i}$ is the n_i *x 1* vector of residuals in environment *i.* Covariance between different random effects in the model within an environment was assumed nil.

$E(y_i)=1_i \mu_i+\mathbf{X}_i \beta_i$, $E(g_i)=\mathbf{0}$, $E(g_{mi})=\mathbf{0}$, and $E(e_i)=\mathbf{0}$. 4-14

$$Var\begin{bmatrix} g_1 \\ g_2 \end{bmatrix} = \frac{1}{4}\begin{bmatrix} \sigma_{a1}^2 & \sigma_{a12} \\ \sigma_{a12} & \sigma_{a2}^2 \end{bmatrix}, \quad \text{and} \quad Var\begin{bmatrix} e_1 \\ e_2 \end{bmatrix} = \begin{bmatrix} \sigma_{e1}^2 & 0 \\ 0 & \sigma_{e2}^2 \end{bmatrix}. \quad 4\text{-}15$$

The phenotypic value of each individual was determined as the summation of all the pertinent genetic and environmental effects in the models. The levels of each effect were assumed to be a random sample from a large population. Independent standard normal deviates ($\mu=0$, $\sigma^2=1$) were created using SAS Rannor function (SAS® Institute Inc. 1990) to reflect random variation within each effect. The magnitude of variation within each effect was determined by its designed variance. For correlated effects between environment **x** and **y**, the method of Van Vleck (1993) was used, in which $E_x=x\sigma_x$, and $E_y=x\frac{\sigma_{xy}}{\sigma_x}+y\sqrt{\sigma_y^2-(\frac{\sigma_{xy}}{\sigma_x})^2}$, where E_x and E_y are the correlated effects in environment **x** and **y**, x and *y* are independent vectors of standard normal deviates, and σ_x^2, σ_{xy} and σ_y^2 are the designed variance and covariance for E_x and E_y respectively. Since estimates of type B genetic correlations are invariant with respect to data standardization and the phenotypic variance in each environment can always be standardized to one, heterogeneous variances between two environments can thus be reflected by their difference in heritability. In this study, several combinations of heritabilities between two environments were considered to represent the situations of heterogeneous variances (Table 4-1 to 4-3).

Block effects can be major sources of variation that contribute to differences among genetic group means when data are unbalanced. This is because block effects are confounded with genetic effects due to non-orthogonality. The assumption of fixed block effects can help

remove these confounding effects. Therefore, block effects were intentionally treated as fixed effects in this study. However, for convenience in data generation, blocks were created as if they were random samples from a large normally distributed population (μ_β=0, σ_β^2 =2), and that the observed variation among blocks was twice as large as the phenotypic variation within a block (i.e., σ_β^2 =2).

After balanced data were generated, 30% mortality was simulated to all data samples by randomly deleting observations. For each set of designed genetic parameters between two environments (h_1^2, h_2^2 and r_B), 500 independent simulation runs were performed.

Methods of comparison

Type B genetic correlations were estimated for each data sample from four different estimating formulas: Yamada method I (Eq.4-1), Burdon method (Eq.4-2); the GCA approach proposed in this study; and an alternative Yamada's formula II, which has often been used in the form:

$$\hat{r}_B = \frac{\hat{\sigma}_g^2}{\hat{\sigma}_g^2 + \hat{\sigma}_I^2}, \qquad 4\text{-}16$$

where the definition of each element is the same as in Eq.4-1.

Two main criteria were used for evaluating the 4 estimation methods. First, empirical bias was calculated as the difference between means of estimated and true type B genetic correlations over 500 random samples for a given population, i.e., $Bias = \bar{r}_{\hat{B}} - \bar{r}_{tB}$, where, $\bar{r}_{\hat{B}} = \frac{1}{N}\sum_i^N \hat{r}_{B_i}$ with $\hat{r}_{B_i}$ being the estimated type B genetic correlation of the ith sample; $\bar{r}_{tB} = \frac{1}{N}\sum_i^N r_{tB_i}$ with r_{tBi} being the true type B genetic correlation of the ith sample. N is the total number of random samples.

Analysis of variance (ANOVA) was conducted for each level of the genetic parameters to detect the statistical significance between the estimated and true type B genetic correlations. A LSD multiple comparison were consequently performed after ANOVA revealing significant differences. The biases showing significant difference from zero were marked.

The second criteria used in the evaluation was the mean-distance (MD) between the estimated and true type B genetic correlations which was calculated as: $MD=\frac{1}{N}\sum_{i}^{N}|\hat{r}_{B_i}-r_{tB_i}|$.

Like for bias, analysis of variance and multiple comparison (LSD) were performed to detect statistical differences among the four estimation methods under each level of genetic parameter. For each level of genetic parameters, the estimation precision of Yamada II was used as comparison standard and estimation precisions of Yamada I, Burdon, and the GCA approach were identified if they were significantly higher or lower than that of Yamada II.

Due to sampling errors, the estimates of variance components may result in the estimatesof type B genetic correlation beyond parameter space (i.e., $\hat{r}_{B_i}<0$ or $\hat{r}_{B_i}>1.0$). In these cases, estimated correlations were modified to the theoretical boundary value (0 or 1.0) following Singh et al. (1997).

Results of Numerical Comparisons

Bias

When data were balanced, all methods except for Yamada II yielded identical estimates even though genetic and environmental variances were heterogeneous (data not shown). When data were balanced within an environment but were unbalanced between

environments in terms of the number of replications (blocks), Burdon's method (Eq.4-2) and the GCA approach (Eq.4-12) yielded identical estimates of type B genetic correlations, which generally had smaller empirical biases than the Yamada methods (Table 4-1). The larger magnitude of biases from Yamada's formulae (I and II) were related to the differences in heritabilities and the number of replications between two environments. The larger the difference between two environments in these factors, the greater the biases from Yamada's formulae.

When data were unbalanced within an environment due to missing values caused by mortality (30%), but approximately equally replicated across environments in terms of the number of blocks, Burdon's formula (Eq.4-2) generally yielded the largest downward biases. In contrast, for such data structure, Yamada's methods yield almost unbiased estimates. When data were unbalanced both within and between environments, the GCA approach consistently yielded the smallest biases which were, for most cases, not significantly different from zero (Table 4-1).

Estimation precision

The precision of estimates of type B genetic correlation is represented by the mean-distance (MD) between estimated and true type B genetic correlations. The MDs were affected by both the heritabilities in two environments and the estimating methods when data were unbalanced (Table 4-2). For all methods, the higher the heritability in two environments, the smaller the MD. When the lower heritability between two environments was accompanied by a larger number of replications, Burdon's methods (Eq.4-2) and the GCA approach tended to yield smaller MD. Whereas, when the lower heritability between

Table 4-1. Empirical biases of type B genetic correlations from four estimating methods under different genetic architectures in two environments as represented by their narrow sense heritabilities (h_1^2 and h_2^2) and true type B genetic correlations. Experimental design is randomized complete block design with single-tree plots in both environments.

Env. 1	Env. 1	True	Env. 1	Env. 1	Mortality	Empirical Biases of r_B estimates			
h_1^2	h_2^2	r_B	b_1	b_2	%	Yamada I	Yamada II	Burdon	GCA Approach
0.2	0.5	0.75	40	20	0	0.064*	-0.031*	0.015	0.015
			40	20	30	0.068*	-0.025	-0.134*	0.007
			20	40	0	-0.005	-0.082*	0.014*	0.014
			20	40	30	-0.059*	-0.140*	-0.166*	-0.014
			40	10	0	0.142*	0.050*	0.021	0.021
			10	40	0	-0.068*	-0.123*	0.005	0.005
0.2	0.4	0.75	40	40	30	0.000	-0.057*	-0.135*	0.014
			40	20	0	0.052*	-0.006	0.014	0.014
			40	20	30	0.053*	-0.002	-0.153*	0.009
			20	40	0	-0.010	-0.063*	0.020	0.020
			20	40	30	-0.066*	-0.117*	-0.180*	0.020
			40	10	0	0.104*	0.057*	0.024	0.024
			10	40	0	-0.068*	-0.123*	0.005	0.005
0.2	0.3	0.75	20	20	30	0.029	-0.002	-0.180*	0.023
			20	10	0	0.052*	0.022	0.013	0.013
			20	10	30	-0.014	-0.050*	-0.212*	0.001
			10	20	0	-0.015	-0.052*	0.005	0.005
			10	20	30	-0.022	-0.064*	-0.212*	-0.031*
0.2	0.1	0.75	40	4	30	0.000	-0.053*	-0.161*	-0.005
			40	20	0	0.074*	0.026	0.035*	0.035*
			20	10	30	-0.013	-0.003	-0.183*	-0.012
			10	20	0	-0.069*	-0.089*	-0.014	-0.014
			20	40	30	-0.043*	-0.032*	-0.165*	-0.019
0.1	0.4	0.60	20	20	30	0.010	-0.115*	-0.122*	0.008
			20	10	0	0.153*	-0.009	0.020	0.020
			20	10	30	0.111*	0.017	-0.156*	0.021
			10	20	0	0.007	-0.116*	0.030*	0.030*
			10	20	30	0.055*	-0.181*	-0.193*	0.073*

Note: Biases are calculated based on 500 simulated random samples for each combination of genetic architecture and data imbalance; b1 and b2 are the numbers of blocks in environment 1 and 2. *Bias is significant different from zero at the probability level of $\alpha \leq 0.05$.

Table 4-2. Mean-distance (MD) between estimated and true type B genetic correlations from four estimation methods under different genetic architectures in two environments (as represented by their narrow sense heritabilities (h_1^2 and h_2^2) and true type B genetic correlations). The experimental design is randomized complete block with single-tree plots in both environments.

Env. 1 h_1^2	Env. 2 h_2^2	True r_B	Env. 1 b_1	Env. 2 b_2	Mortality %	Mean-distance			
						Yamada I	Yamada II	Burdon	GCA Approach
0.2	0.5	0.75	40	20	0	0.108*+	0.095	0.086*-	0.086*-
			40	20	30	0.137	0.123	0.177*+	0.117
			20	40	0	0.101	0.110	0.111	0.111
			20	40	30	0.143*-	0.165	0.213*+	0.144*-
			40	10	0	0.173*+	0.127	0.113*-	0.113*-
			10	40	0	0.157	0.158	0.174	0.174
0.2	0.4	0.75	40	40	30	0.175	0.167	0.248*+	0.175
			40	20	0	0.107	0.096	0.093	0.093
			40	20	30	0.138	0.127	0.198*+	0.125
			20	40	0	0.106	0.112	0.113	0.113
			20	40	30	0.151	0.161	0.235*+	0.148
			40	10	0	0.157	0.136	0.128*-	0.128*-
			10	40	0	0.164	0.154	0.173*+	0.173*+
0.2	0.3	0.75	20	20	30	0.132	0.128	0.223*+	0.131
			20	10	0	0.141	0.136	0.135	0.135
			20	10	30	0.140	0.138	0.253*+	0.179*+
			10	20	0	0.137	0.135	0.135	0.135
			10	20	30	0.183	0.181	0.292*+	0.199
0.2	0.1	0.75	40	40	30	0.175	0.167	0.248*+	0.175
			40	20	0	0.189	0.199	0.169*-	0.169*-
			40	20	30	0.211*-	0.206	0.294*+	0.213
			20	40	0	0.201	0.174	0.206*+	0.206*+
			20	40	30	0.256	0.202	0.320*+	0.267*+
0.1	0.4	0.60	20	20	30	0.195	0.180	0.237*+	0.198
			20	10	0	0.168	0.164	0.166	0.166
			20	10	30	0.291*+	0.214	0.253*+	0.222
			10	20	0	0.184	0.194	0.221	0.221
			10	20	30	0.231	0.229	0.324*+	0.274

Note: Mean-distance is calculated based on 500 simulated random samples for each combination of genetic architecture and data imbalance. b_1 and b_2 are the number of blocks in individual environment. [1] b1 and b2 are the number of blocks in individual environment. [2] - and [3] + indicate that MD calculated from Yamada I, Burdon or the GCA-approach is significantly smaller (better) or larger (worse) than MD calculated from Yamada II at the probability level $\alpha \leq 0.05$.

two environments was accompanied by fewer number of replications, Yamada's method I & II (Eq.4-1 & Eq.4-16) resulted in smaller MD. Consistently, Burdon's method yielded considerably larger MD than the other methods when 30% mortality was present in the experimental data. For all cases, the GCA approach had MDs less than or equal to Burdon's method, depending on data imbalance (Table 4-2).

When the estimation precision of Yamada II was used as a standard for comparison, the precision of Yamada I or the GCA-approach differed significantly only for a few levels of genetic parameters, in which they were either significantly higher (MD-) or lower (MD+) than that of Yamada II (Table 4-2). For the majority of levels of genetic parameters, however, these three estimation methods showed no significant differences. For the Burdon's method, the estimation precision was almost always significantly lower than that of Yamada II when 30% of mortality was present in the data.

Out-of-bound estimates

Estimates of type B genetic correlation from the simulated data sets often fell outside the theoretical parameter space of 0 to 1 (Burdon 1977). All methods except for Yamada II yielded such results, especially when the true population heritabilities were low (data not shown). When true population heritabilities were 0.1 to 0.2 in the simulated data, up to 25% of estimates of type B genetic correlations were either greater than 1 or less than zero even though data were balanced. It was observed that the out-of-bound estimates of type B genetic correlation were mainly caused by near-to-zero estimates of genetic variance(s) from one or both of the testing environments. Being in the denominator of the estimating equations, this produced irregular values of estimates of type B genetic correlation. Figure 4-1 indicated the

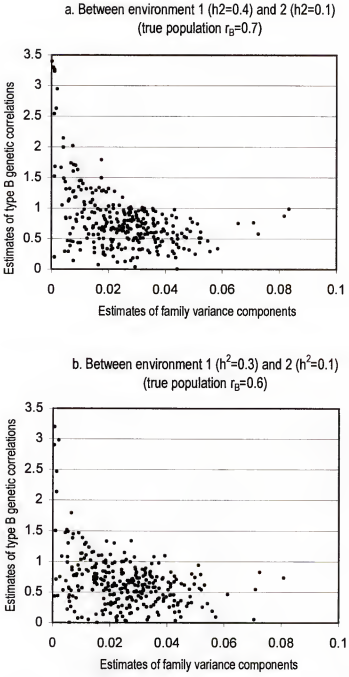

Figure 4 –1. Effects of genetic variance component estimates on the number of out-of-bound estimates of type B genetic correlation

effects of near to zero estimates of family variance components on the number of out-of-bound estimates of type B genetic correlations. The number of out-of-bound estimates became steadily smaller when the true population heritabilities in two environments were higher (>0.2). There was, however, no significant difference among Yamada I, Burdon and the GCA approach in producing out-of-bound estimates for a given population genetic structure.

Correlation between true and estimated type B genetic correlations

Pearson correlation coefficients between the true and estimated genetic correlations are given in Table 4-3. Each Pearson correlation coefficient was calculated based on 500 random data samples each having a pair of true and estimated type B genetic correlations from a given estimation method under a given level of genetic parameters. A higher correlation between the true and estimated type B genetic correlation indicates a better correspondence between them and, therefore, a higher quality for the estimation method. In this simulation study, the ANOVA based methods (i.e., Yamada I and II) showed detectable differences from the correlation-based methods (i.e., the Burdon and GCA approaches). For data with fewer replications in environments of low heritability, the former seemed to be better, whereas for data with fewer replications in environments with higher heritability, the latter seemed to be better. In general, the GCA approach showed consistent improvement over Burdon's approach, and Yamada II performed better than Yamada I.

Table 4-3. Correlation between estimated and true type B genetic correlations from four estimating methods under different genetic architectures in two environments as represented by their narrow sense heritabilities (h_1^2 and h_2^2) and true type B genetic correlations. Experimental design is randomized complete block design with single-tree plots in both environments.

Env. 1 h_1^2	Env. 1 h_2^2	True r_B	Env. 1 b_1	Env. 2 b_2	Mortality %	Correlation Coefficients			
						Yamada I	Yamada II	Burdon	BV Approach
0.2	0.5	0.75	40	20	0	0.602	0.600	0.642	0.642
			40	20	30	0.482	0.501	0.431	0.533
			20	40	0	0.581	0.590	0.546	0.546
			20	40	30	0.420	0.454	0.370	0.394
			40	10	0	0.502	0.547	0.578	0.578
			10	40	0	0.489	0.438	0.387	0.387
0.2	0.4	0.75	40	40	30	0.522	0.510	0.411	0.520
			40	20	0	0.498	0.470	0.518	0.518
			40	20	30	0.397	0.413	0.352	0.434
			20	40	0	0.458	0.459	0.433	0.433
			20	40	30	0.304	0.334	0.216	0.292
			40	10	0	0.432	0.465	0.479	0.479
			10	40	0	0.354	0.395	0.319	0.319
0.2	0.3	0.75	20	20	30	0.356	0.378	0.237	0.359
			20	10	0	0.329	0.328	0.328	0.328
			20	10	30	0.203	0.247	0.086	0.123
			10	20	0	0.315	0.349	0.319	0.319
			10	20	30	0.123	0.138	0.076	0.168
0.2	0.1	0.75	40	40	30	0.273	0.312	0.079	0.275
			40	20	0	0.310	0.322	0.345	0.345
			40	20	30	0.301	0.306	0.192	0.269
			20	40	0	0.286	0.333	0.278	0.278
			20	40	30	0.220	0.231	0.142	0.212
0.1	0.4	0.60	20	20	30	0.267	0.265	0.144	0.272
			20	10	0	0.417	0.438	0.466	0.466
			20	10	30	0.237	0.248	0.259	0.280
			10	20	0	0.373	0.412	0.271	0.278
			10	20	30	0.226	0.258	0.089	0.187

Note: b_1 and b_2 are the numbers of blocks in environment 1 and 2.

Discussion

The Yamada and Burdon methods of type B genetic correlation have been routinely used in forest genetic data analysis, but the properties of the estimates are not well understood. Simulation studies with data from populations of known genetic parameters provided useful information about the desirable and undesirable properties of estimates from these methods.

The new GCA approach using predicted parental GCA effects with the technique of BLUP has shown considerable improvement over Burdon's method in producing smaller empirical bias, smaller mean distance and an improved relationship between the true and estimated type B genetic correlations when data are unbalanced and have heterogeneous variances. This is consistent with theoretical considerations as discussed in this chapter.

For the Yamada (1962) methods (I & II), the absolute values of empirical biases were not numerically large although most of them were statistically different from zero. This was especially true when data samples were not extremely unbalanced between two environments and the heterogeneity of variances was not very large. Only in a few cases where data were extremely unbalanced between two environments in terms of their relative sizes (i.e., the numbers of blocks) and the heterogeneity was large ($h_1^2/h_2^2 > 2$), were some larger absolute values of biases observed. Compared with the large sampling errors of the estimates of type B genetic correlations with low heritabilities (Table 4-2), the magnitudes of biases may be of little practical significance. In this sense, the results may suggest that Yamada's methods can be reasonably used in most cases of forest genetic tests that involves moderately

unbalanced data, even though the Yamada's methods have some theoretical inadequacies (Fernando et al. 1984).

The approach of Yamada II (Eq.4-16) was the only univariate method that consistently yielded estimates of type B genetic correlation within the theoretical parameter boundaries (0 to 1) when estimates of negative variances were accepted as zero. Due to the fact that the effect of sampling error of variance components on the estimation of type B genetic correlation cannot be effectively controlled by the other three estimating methods, Eq.4-16 (Yamada II) was the most empirically robust univariate method in estimating type B genetic correlation. Although this approach is not theoretically well defined for data with heterogeneous variances (Yamada 1962), it did produce empirically equal or more desirable properties of the estimates than the Yamada I (Eq.4-1) and Burdon's. As long as two environments have small differences in the numbers of replications and heritabilities, the empirical bias from this approach is minimal and may be negligible (Table 4-1) after data standardization removing scale effects. Considering its other benefits, such as the computational convenience and satisfying the practical needs of maintaining estimates of genetic correlations within theoretical parameter space for the purpose of indirect selection, this method would be empirically more suitable than Yamada I and Burdon approaches. If heterogeneous variances among environments are extremely large, however, this method is known to yield severe biases (Dutilleul and Carrière 1998).

Burdon's formula is sensitive to mortality if block effects are large. This result may partially be attributable to the relatively large block effects simulated in this study. Large block effects are common, however, in forest genetic tests given the heterogeneous testing

environments of forest lands. Like the GCA approach, Burdon's method is robust to unequal data sizes and heterogeneous variances between two environments as long as mortality is low within a progeny test. But when mortality is high and block effects are large, this method generally produces estimates with large downward biases and significantly lower estimation precision, which are inferior to other methods.

For convenience, the derivation of the GCA approach and the simulations were based on half-sib families. The method, however, may be extendable to estimate additive type B genetic correlations with full-sib families. For full-sib families, the predicted parental GCA effects, not the full-sib family means, are used in Eq.4-12 and the estimates of type B genetic correlations should be additive. It should be noted that for full-sib family, estimates of type B genetic correlation from Burdon's method are not additive genetic correlations but genotypic correlations because additive genetic effects and dominance effects are not separated in that method. Since estimates of additive type B genetic correlations are needed to predict genetic gains from indirect selection (Falconer 1989), further studies may be necessary to examine estimation methods for their ability to estimate additive type B genetic correlations with full-sib genetic test data.

Like the methods of Yamada I and Burdon, the GCA approach only estimates type B genetic correlations between pairs of environments. To get the averaged estimate of type genetic correlation among multiple environments, one may either take the average of paired estimates (Dieters et al. 1995) or pool the data of appropriate environments and predict parental GCA effects with more complex mixed linear models. In this regard, the GCA approach is not computational efficient as compared with Yamada II. But when the

assumption of homogeneous variances cannot be made or the experimental designs are different among experiments, it can be a useful alternative.

Conclusion

The GCA approach of type B genetic correlation estimation has provided an useful tool in handling unbalanced data with heterogeneous variances. This method is relatively easy to use when BLUP computer packages are available. since the GCA approach calculates type B genetic correlations after properly removing fixed effects and adjusting for heterogeneous variances, estimates of type B genetic correlations generally have better or equal properties in terms of unbiasedness and precision as compared with other methods. For experimental data composed of different field experimental designs, severe data imbalances and large heterogeneous variances, the GCA approach is a viable option for estimating type B genetic correlations. Yamada's (1962) methods may be appropriate to estimate type B genetic correlation with acceptable or negligible biases for slightly or moderately unbalanced forest genetic test data which have moderately heterogeneous variances.

CHAPTER 5
COMPARISON OF MULTIVARIATE AND UNIVARIATE METHODS FOR ESTIMATING TYPE B GENETIC CORRELATIONS

Introduction

In quantitative forest genetic data analysis, type B genetic correlations (Burdon 1977) have been routinely estimated using the univariate methods of Yamada (1962) and Burdon (1977) (Burdon 1977; Johnson and Burdon 1990; Woolaston et al. 1991; Hodge and White 1992; Hodge and Purnell 1993; Adams et al. 1994; Dieters et al. 1995; Dieters 1996; Pswarayi et al. 1997). For pairs of genetic tests, these methods first estimate genetic variances and covariances using univariate linear models in one or/and two environments separately and then calculate genetic correlations according to the described procedures (Yamada 1962; Burdon 1977). Although these univariate methods provide considerable flexibility in minimizing computational demands and facilitating suitable computer softerwares (such as SAS®), arguments suggest that estimates of genetic correlations from these univariate methods are less satisfactory for some data structures (Fernando et al. 1984; Dutilleul and Carrière 1998).

One of the undesirable aspects of these univariate methods is their inability to yield unbiased estimates of type B genetic correlations for unbalanced experimental data accompanied with heterogeneous variances across environments. For example, theoretical

considerations and empirical evidence have suggested that serious biases are associated with estimates of type B genetic correlation from Yamada's (1962) methods when data are unbalanced in terms of experiment sizes and when heterogeneous variances are present across environments (Fernando et al. 1984; Ito and Yamada 1990; Dutilleul and Carrière 1998). Severe biases for estimates of type B genetic correlations were also found with Burdon's method when mortalities cause data imbalance within a genetic test (Chapter 4 of this dissertation). Although improvement to univariate methods can be made (Dutilleul and Carrière 1998; Chapter 4 of this dissertation) so that unbiased estimates of type B genetic correlations are still obtainable for unbalanced data with heterogeneous variances, some of the procedures become more computationally complex and inconvenient.

Another concern with univariate methods is that type B genetic correlation estimates are often out of the theoretical parameter space and results are, therefore, difficult to apply in practical applications. For instance, both empirical evidence (Koots and Gibson 1996) and simulation study (chapter 4 of this dissertation) have indicated that the frequency of out-of-bound estimates increases with the decrease of true population heritabilities. Out-of-bound estimates often occur when near-to-zero estimates of genetic variance are obtained in one or two of the concerned environments, while the estimates of genetic covariance are relatively large. Potential reasons for out-of-bounds estimates are primarily attributable to sampling errors of genetic variances and covariances, but may also be related to the fact that genetic variance-covariances are not estimated from a closed system with these univariate methods.

A third potential question facing univariate methods of type B genetic correlation estimation is that genetic relatedness and inbreeding among and within genetic groups cannot

be properly accounted for. Since the possibility of genetic relatedness among genetic groups (especially among full-sib families) increases as tree improvement programs progress into advanced generations (White et al. 1993; Borralho and Dutkowski 1998), the assumption with univariate statistical methods that genetic groups are independent random samples of a large population would be violated if genetic relatedness does exist. Failure to account for genetic relatedness and inbreeding may cause inaccurate estimates of genetic variance and covariances and potentially result in biased estimates of type B genetic correlations.

Multivariate methods can estimate genetic variances and covariances simultaneously using restricted maximum likelihood (REML) approach with an iterative procedure (Patterson and Thompson 1971; Schaeffer and Wilton 1978). For multivariate methods, measurements from different environments are treated as different traits with different variance and covariance structures. Consequently, the problem of heterogeneous variances facing univariate methods is properly solved. It is believed that REML approach is generally more desirable than ANOVA (analysis of variance) methods in handling with unbalanced data for the purpose of variance component estimation (Searle et al. 1992; Huber et al. 1994). In addition, some multivariate methods can apply constraints to estimates of genetic variances and covariances so that estimates of genetic correlations stay within the theoretical parameter space (Boldman et al. 1995). Furthermore, multivariate REML methods can make use of pedigree information so that genetic relatedness among genetic groups is properly treated in the process of variance component estimation (Boldman et al. 1995; Gilmour et al.1997).

Despite the potential advantageous properties of multivariate methods in estimating type B genetic correlations, uncertainty remains as to whether constrained multivariate procedures yield unbiased estimates. It is also unclear how many environments should be used in a constrained system to enhance the quality of estimates. The purpose of this study was to numerically compare some multivariate and commonly used univariate methods for type B genetic correlation estimation using simulated forest genetic data. Specifically, we examined these estimation methods in terms of unbiasedness and precision, as well as the distributional properties of the estimates under different genetic parameters.

Material and Methods

Data Generation

Simulated data were used in numerical comparisons because for each simulated data set the true underlying type B genetic correlation is known and can used to evaluate the qualities of estimates. Data were generated based on a randomized complete block (RCB) design with one tree per family per plot (i.e., single-tree plots) which is recommended by several studies in forest genetic testing (Lambeth and Gladstone 1983; Loo-Dinkins and Tauer 1987; Loo-Dinkins et al. 1990; White 1996). Genetic structures of the data were simulated based on half-sib families created from a polymix mating design with 120 female parents.

In the field experimental designs, it was assumed that these 120 half-sib families were tested over 4 environments, each having 90 families and 20 blocks. It was further assumed that there were 60 half-sib families in common for any paired progeny tests, but there was

no family in common across the 4 environments. The linear model used in data generation across 4 testing environments is given in matrix notation as:

$$y=\begin{bmatrix} y_1 \\ \vdots \\ y_4 \end{bmatrix}=\begin{bmatrix} 1_1\mu_1 \\ \vdots \\ 1_4\mu_4 \end{bmatrix}+\begin{bmatrix} X_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & X_4 \end{bmatrix}\begin{bmatrix} \beta_1 \\ \vdots \\ \beta_4 \end{bmatrix}+\begin{bmatrix} Z_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & Z_4 \end{bmatrix}\begin{bmatrix} g_1 \\ \vdots \\ g_4 \end{bmatrix}+\begin{bmatrix} Z_{m1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & Z_{m4} \end{bmatrix}\begin{bmatrix} g_{m1} \\ \vdots \\ g_{m4} \end{bmatrix}\begin{bmatrix} e_1 \\ \vdots \\ e_4 \end{bmatrix} \quad 5\text{-}1$$

where $\boldsymbol{y_i}$ is a n_i *x 1* vector of phenotypic observations in environment *i, i=1,...4*; n_i is the number of observations in environment *i*; μ_i is the overall mean in environment *i* and 1_i is an n_i *x 1* vector of 1s; $\mathbf{X_i}$ is the incidence matrix relating to block effects (vector $\boldsymbol{\beta_i}$) in environment *i;* $\mathbf{Z_i}$ is the incidence matrix relating to female parental genetic effects (vector $\boldsymbol{g_i}$) in environment *i;* $\mathbf{Z_{mi}}$ is the incidence matrix relating to the genetic effects of Mendelian sampling (vector $\boldsymbol{g_{mi}}$) in environment *i*; $\boldsymbol{e_i}$ is the n_i *x 1* vector of residuals in environment *i.* Covariance between random effects (i.e., female parent and residual) in the model was assumed nil, such that

$$E(y_i)=1_i\,\mu_i+\mathbf{X_i}\,\beta_i, \qquad E(g_i)=\mathbf{0}, \quad E(g_{mi})=\mathbf{0}, \quad \text{and} \quad E(e_i)=\mathbf{0};$$

$$G=Var\begin{bmatrix} g_1 \\ g_2 \\ g_3 \\ g_4 \end{bmatrix}=\frac{1}{4}\begin{bmatrix} \sigma^2_{a1} & \sigma_{a12} & \sigma_{a13} & \sigma_{a14} \\ \sigma_{a21} & \sigma^2_{a2} & \sigma_{a23} & \sigma_{a24} \\ \sigma_{a31} & \sigma_{a32} & \sigma^2_{a3} & \sigma_{a34} \\ \sigma_{a41} & \sigma_{a42} & \sigma_{a43} & \sigma^2_{a4} \end{bmatrix} \quad R=Var\begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \end{bmatrix}=\begin{bmatrix} \sigma^2_{e1} & 0 & 0 & 0 \\ 0 & \sigma^2_{e2} & 0 & 0 \\ 0 & 0 & \sigma^2_{e3} & 0 \\ 0 & 0 & 0 & \sigma^2_{e4} \end{bmatrix}.$$

The phenotypic value of each individual was determined as the summation of all genetic and environmental effects in the model. The levels of each effect were assumed to be a random sample from a large normal population. Independent standard normal deviates (μ=0, σ^2 =1) were created using the SAS Rannor function (SAS® Institute Inc. 1990) to reflect random variation within each effect. The magnitude of variation within each effect

was determined by its designed variance. Correlated female additive genetic effects among the 4 testing environments were created as:

$$\mathbf{A} = \mathbf{B}'\mathbf{C} \qquad 5\text{-}2$$

where **A** is the matrix of additive genetic effects of female parents, **B** is a matrix of the square root (Cholesky decomposition) of designed genetic variance-covariance matrix **G**, and **C** is a column vector of independent standard normal random deviates, such that

$$\text{Var}(\mathbf{A}) = \mathbf{B}'\text{Var}(\mathbf{C})\mathbf{B} = \mathbf{B}'\mathbf{B} = \mathbf{G} \qquad 5\text{-}3$$

Heterogeneous genetic variances among progeny tests are reflected by the genetic variances and covariances in matrix **G**. The designed population genetic parameters such as heritability and type B genetic correlations were intentionally simulated to have a relatively large variation among the 4 environments aimed to represent a wide range of situations that may exist among real forest genetic data sets (Table 5-1). Without losing generality, phenotypic variance within each progeny test was set to 1.0 due to the fact that data standardization is highly recommended in forest genetic data analysis (Wu 1993; White 1996; Dieters 1996) to remove scale effects and that data standardization can always adjust phenotypic variance to 1 within a single environment.

Block effects were treated as fixed effects in this study for three reasons: (1) variance component estimation for random female genetic effect is not affected under customary linear models (i.e., no covariance assumed between random effects) whether block is treated as random or fixed effect; (2) treating block as fixed effect facilitates the application of multivariate methods in this study; and (3) the assumption of fixed block effects can help remove block effects from confounding genetic effects when data are unbalanced. However,

for convenience in data generation, levels of block were created with the assumption that they were random samples from a normally distributed population (μ_b=0, σ_b^2 =2), and that the observed variation among blocks was twice as large as the phenotypic variation within a block (i.e., σ_b^2 =2).

After data were generated, 30% mortality was randomly simulated in all progeny tests by random deletion. A total of 300 independent simulated data sets, each containing 4 environments and 20 blocks, were generated and analyzed in this study.

Table 5-1. Designed heritabilities and type B genetic correlations for an arbitrary continuous trait among four simulated environments.

	Environment			
	1	2	3	4
	Heritability and type B genetic correlations			
1	0.40	0.90	0.80	0.70
2		0.30	0.70	0.60
3			0.20	0.50
4				0.10
	Designed genetic variance and covariance matrix (G)			
1	0.1000	0.0779	0.0566	0.0350
2		0.0750	0.0429	0.0260
3			0.0500	0.0177
4				0.0250

Shaded areas are narrow sense heritabilities of the four environments.

Estimation of Type B Genetic Correlations

Multivariate methods

Multivariate computer programs MTDFREML (Boldman et al. 1995) and ASREML (Gilmour et al.1997) were used to analyze each of the 300 data sets by treating the measurements of an arbitrary continuous trait from different environments as different traits. MTDFREML is a computer software which uses a simplex method to approach the convergence for variance component estimation and allows for constraining estimates of genetic correlations within the theoretical parameter space. ASREML is, on the other hand, a computer program which uses an average information algorithm (Gilmour et al. 1997) and sparse matrix techniques to efficiently solve large mixed models; it does not constrain estimates of genetic correlations within theoretical parameter space. Therefore, results from MTDFREML and ASREML were used to represent, respectively, the constrained and unconstrained multivariate estimates for type B genetic correlations.

For both multivariate methods, input data structures were modified (Table 5-2) in order to estimate genetic variances and covariances for the same trait measured in different environments (type B) rather than for different traits measured on the same individuals (type A). Convergence criteria were set for MTDFREML as MVFV (Minimum Variance of Function Values in Simplex) $\leq 10^{-9}$, and for ASREML, $|(-2L_{n+1} - (-2L_{n})| \leq 0.002$, respectively, following the instructions of program manuals (Boldman et al. 1995; Gilmour et al. 1997). True genetic variance-covariance components were used as priors for starting the iterative processes. For both multivariate methods, two different grouping of

environments were used to estimate variance components based on the assumption that estimates of genetic variances and covariances are system dependent so that estimates of type B genetic correlations are not identical if they are estimated from pair-wise or over all sites. The small grouping contained only pairs of environments while the larger grouping contained all four environments and estimated all pair-wise genetic correlations simultaneously.

Table 5-2. Illustration of data structure used in multivariate analysis to estimate type B genetic correlations. Experimental design is assumed as a randomized complete block design with 3 environments, each having 3 blocks with one tree per family per block. Observations from different environments are treated as different traits.

Environment	Block	Family	Trait 1	Trait 2	Trait 3
1	1	1	10.51	.	.
1	1	2	9.83	.	.
1	1	3	7.78	.	.
1	2	1	8.39	.	.
1	2	2	7.67	.	.
1	2	3	6.78	.	.
1	3	1	12.34	.	.
1	3	2	11.23	.	.
1	3	3	10.98	.	.
2	4	1	.	8.65	.
2	4	2	.	8.21	.
2	4	3	.	7.67	.
2	5	1	.	9.69	.
2	5	2	.	8.76	.
2	5	3	.	8.65	.
2	6	1	.	6.67	.
2	6	2	.	7.43	.
2	6	3	.	5.89	.
3	7	1	.	.	12.34
3	7	2	.	.	13.45
3	7	3	.	.	10.56
3	8	1	.	.	13.45
3	8	2	.	.	12.56
3	8	3	.	.	11.98
3	9	1	.	.	12.17
3	9	2	.	.	13.64
3	9	3	.	.	15.48

Note that blocks 1-3 are from environment 1, blocks 4-6 are from environment 2 and blocks 7-9 are from environment 3. Dots stand for missing values, which are necessary and intentionally given in this data structure.

Univariate methods

Univariate methods used in this study included the traditional methods of Yamada (1962), Burdon (1977) and a GCA approach (Chapter 4 of this dissertation). A previous comparative simulation study suggested that the GCA approach generally yields more desirable properties of the estimates of type B genetic correlation when data are severely unbalanced and heterogeneous variances exist among environments (Chapter 4 of this dissertation).

For the Yamada method, the type B genetic correlation is estimated as:

$$\hat{r}_B = \frac{\hat{\sigma}_g^2}{\hat{\sigma}_g^2 + \hat{\sigma}_I^2 - \frac{(\hat{\sigma}_{g1} - \hat{\sigma}_{g2})^2}{2}} \qquad 5\text{-}4$$

where $\hat{r}_B$ is the estimated type-B genetic correlation, $\hat{\sigma}_g^2$ is the genetic variance component estimated from a two-way analysis of variance involving data from two environments assuming homogeneous variance between them, $\hat{\sigma}_I^2$ is the estimate of variance component for the effect of G x E interaction, $\hat{\sigma}_{g1}^2$ and $\hat{\sigma}_{g2}^2$ are, respectively, the estimates of genetic variance components within environment 1 and 2. Often used in forest genetic studies is a simplified formula of Eq.4 which is:

$$\hat{r}_B = \frac{\hat{\sigma}_g^2}{\hat{\sigma}_g^2 + \hat{\sigma}_I^2} \qquad 5\text{-}5$$

where elements in Eq.5 are the same as in Eq.4. For convenience, we refer Eq.4 to Yamada I and Eq.5 to Yamada II.

With Burdon's method, the type B genetic correlation is estimated as:

$$\hat{r}_B = \frac{\hat{r}_{xy}}{\hat{h}_{\bar{x}}\hat{h}_{\bar{y}}} \qquad 5\text{-}6$$

where $\hat{r}_{xy}$ is the phenotypic correlation between genetic group (i.e. half-sib families) means in environments **x** and **y**, and $\hat{h}_{\bar{y}}$ and $\hat{h}_{\bar{y}}$ are square-roots of the heritabilities of the genetic group means in environments **x** and **y**, respectively.

For the GCA approach, parental GCA effects are first predicted using the technique of univariate best linear unbiased prediction (BLUP) in each environment and these are adjusted to calculate type B genetic correlation as:

$$\hat{r}_B = \frac{Cov(\frac{\hat{g}_{xi}}{\hat{r}_{axi}}, \frac{\hat{g}_{yi}}{\hat{r}_{ayi}})}{\overline{\hat{r}_{ax}\hat{r}_{ay}}\sqrt{Var(\frac{\hat{g}_{xi}}{\hat{r}_{axi}})Var(\frac{\hat{g}_{yi}}{\hat{r}_{ayi}})}} = \frac{\hat{r}^*}{\overline{\hat{r}_{ax}\hat{r}_{ay}}}, \qquad 5\text{-}7$$

where $\hat{r}_B$ is the estimate of type B genetic correlation, r^* is a Pearson correlation coefficient between adjusted parental GCA predictions in environments **x** and **y** using BLUP, $\hat{g}_{xi}$ and $\hat{g}_{yi}$ are predicted parental GCA effects in environments **x** and **y** respectively, and $\overline{\hat{r}_{ax}\hat{r}_{ay}}$ is the mean products of adjusted 'prediction accuracy' (see Chapter 4 of this dissertation) in the two environments.

Criteria for Comparisons

After the type B genetic correlations were estimated for each pair of environments within each of the 300 simulated data sets using the above univariate and multivariate

methods, three main criteria were used to evaluate the estimation methods. First, empirical bias was calculated as the difference between means of estimated and true type B genetic correlations over 300 random samples for each pair of environments, i.e., $Bias=\bar{r}_{\hat{B}}-\bar{r}_{tB}$, where $\bar{r}_{\hat{B}}=\frac{1}{N}\sum_{i}^{N}\hat{r}_{B_i}$ with $\hat{r}_{B_i}$ being the estimated type B genetic correlation of the ith sample; $\bar{r}_{tB}=\frac{1}{N}\sum_{i}^{N}r_{tB_i}$ with r_{tBi} being the true type B genetic correlation of the ith sample. N(=300) is the total

number of random samples. The statistical differences of the empirical biases from zero were tested by one-way analysis of variance (ANOVA).

The second criterion was the mean-distance (MD) between the estimated and true type-B genetic correlations which was calculated as: $MD=\frac{1}{N}\sum_{i}^{N}|\hat{r}_{B_i}-r_{tB_i}|$. The smaller the MD, the closer the estimates to their true values, and consequently, the higher estimation precision. The third criterion was the simple correlation between the estimates of type B genetic correlations and the true type B genetic correlations. Higher correlation reflects the better response of estimated type B to the changes of true type B genetic correlation and thus shows better quality of the related estimation methods. Outliers were excluded if their distances to the true values exceeded three times the MD.

Results

Bias

Among the univariate methods, which is always applied to pairs of environments, the GCA approach yielded empirically unbiased estimates of type B genetic correlations for all

balanced and unbalanced data sets (Table 5-3). Burdon's method yielded unbiased estimates when data were balanced within an environment but yielded severely biased estimates when data were unbalanced due to missing values. Yamada I (Eq.5-4) also produced nearly unbiased estimates for almost all data sets. Yamada II (Eq.5-5) tended to yield slightly downward biases, and biases for a few environment pairs became severe when ratios of genetic variances in two environments were greater than 2.

Unconstrained multivariate method ASREML and constrained multivariate method MTDFREML both yielded empirically unbiased estimates of type B genetic correlations when either all four environments or only two environments were included in a closed analytical system (Table 5-3). MTDFREML tended to yield slightly downward biases but the magnitudes were nearly negligible. For a given pair of environments, the large and small grouping, however, differed in the estimates of type B genetic correlations which was indicated by the imperfect correlation between estimates from the two analytical systems for a given pair of environments (data not shown). This implied that, with multivariate methods, data within an analytical system are inter-communicated among environments so that information from a third environment can influence the estimates of type B genetic correlation between the other two environments.

Comparison of the best results from univariate methods with those of multivariate methods in terms of bias did not produce obvious differences. The univariate methods, such as the GCA-approach and Yamada I, produced estimates as unbiased as those from multivariate methods for almost all data structures simulated in this study. Other univariate methods, such as Burdon method and Yamada II did, however, yield empirically biased

estimates of type B genetic correlations when data were unbalanced due to missing values or when heterogeneity of variances among environments was large. Such results were consistent with previous theoretical considerations and empirical studies about these univariate methods (Fernando et al. 1984; Ito and Yamada 1990; Dutilleul and Carrière 1998).

Table 5-3. Empirical biases of type B genetic correlation estimates from different estimation methods for simulated half-sib data tested in a randomized complete block experimental design with single-tree plots.

Mortality	Estimation Method	Environment Pairs					
		1-2	1-3	1-4	2-3	2-4	3-4
0%	MTDFREML 1†	-0.011	-0.011	-0.014	0.015	0.017	0.019
	MTDFREML 2	-0.003	0.001	0.015	0.003	0.009	0.015
	ASREML 1	0.005	0.008	0.067*	0.008	0.069*	0.053*
	ASREML 2	0.002	0.009	0.058	0.036	0.047	0.044
	GCA-approach	0.002	0.008	0.031	0.000	0.019	0.009
	Yamada I	0.000	0.008	0.031	0.006	0.023	0.017
	Yamada II	-0.018*	-0.051*	-0.148*	-0.019	-0.082*	-0.034
	Burdon	0.002	0.008	0.031	0.000	0.019	0.009
30%	MTDFREML 1	-0.023*	-0.019	-0.039	0.017	-0.004	0.020
	MTDFREML 2	-0.009	-0.001	-0.012	0.003	-0.036	0.022
	ASREML 1	0.006	0.020	0.110*	0.022	0.074*	0.098*
	ASREML 2	0.006	0.021	0.134*	0.024	0.085*	0.104*
	GCA-approach	0.004	0.022	0.049	0.018	0.047	0.033
	Yamada I	-0.004	0.015	0.047	0.022	0.041	0.038
	Yamada II	-0.027*	-0.053*	-0.153*	-0.013	-0.091*	-0.031
	Burdon	-0.232*	-0.192*	-0.130*	-0.161*	-0.077*	-0.022

True genetic parameters for different environments are given in Table 5-1. † MTDFREML (or ASREML) 1 and 2 refer to the 4-environment and 2-environment grouping, respectively. * Biases are significantly different from zero at the probability level $\alpha \leq 0.05$.

Precision

The mean distances (MD) between the estimated and true type B genetic correlations from different estimation methods showed consistent differences among the estimation methods (Table 5-4). Constrained estimation methods (i.e., MTDFREML and Yamada II) had smaller MDs than those unconstrained methods (i.e., ASREML, Yamada I, Burdon and the GCA approach). For the constrained multivariate method MTDFREML, the 4-environment grouping consistently had smaller MD than 2-environment grouping. But this was not always true for unconstrained multivariate ASREML. Among all estimation methods, MTDFREML with a 4-environment system persistently yielded the smallest MD for a given pair of environments. This was followed by the MTDFREML with two-environment system, Yamada II, and then the GCA approach and ASREML. The difference between constrained methods and unconstrained methods in MD became even larger when heritabilities in two environments were lowered (Table 5-4).

Regardless of estimation methods, the mean distance between estimated and true type B genetic correlations became steadily larger when heritabilities in two environments were lowered (Table 5-4). For example, for all methods except Burdon's, MDs were not greater than 0.1 between environments 1 and 2, which had heritabilities of 0.4 and 0.3, respectively. In contrast, MDs were greater than 0.2 for almost all methods between environments 3 and 4, which had true heritabilities of only 0.2 and 0.1, respectively.

Table 5-4. Mean-distance (MD) between estimates of type B genetic correlation and their true values for different estimation methods using simulated half-sib families tested in a randomized complete block experimental design with single-tree plots.

Mortality	Estimation Method	Environment Pairs					
		1-2	1-3	1-4	2-3	2-4	3-4
0%	MTDFREML 1	0.060	0.092	0.137	0.119	0.150	0.187
	MTDFREML 2	0.067	0.100	0.166	0.124	0.188	0.205
	ASREML 1	0.076	0.108	0.217	0.126	0.226*†	0.232*†
	ASREML 2	0.080	0.108	0.209	0.126	0.205	0.218
	GCA-approach	0.077	0.106	0.192	0.124	0.197	0.210
	Yamada I	0.071	0.108	0.193	0.126	0.193	0.208
	Yamada II	0.069	0.104	0.176	0.122	0.166	0.180
	Burdon	0.077	0.106	0.192	0.124	0.197	0.210
30%	MTDFREML 1	0.069	0.107	0.176	0.132	0.184	0.230
	MTDFREML 2	0.081	0.125	0.212	0.124	0.200	0.293*†
	ASREML 1	0.099	0.146	0.331*†	0.157	0.283*†	0.306*†
	ASREML 2	0.096	0.148	0.350*†	0.157	0.282*†	0.316*†
	GCA-approach	0.100	0.149	0.316*†	0.156	0.300*†	0.332*†
	Yamada I	0.087	0.142	0.304*†	0.155	0.279*†	0.303*†
	Yamada II	0.084	0.129	0.218	0.144	0.207	0.239
	Burdon	0.245*†	0.255*†	0.305*†	0.239*†	0.313*†	0.293*†

True genetic parameters for different environments are given in Table 5-1. Out-of-bound estimates were accepted with their original values. † MTDFREML (or ASREML) 1 and 2 refer to the 4-environment and 2-environment grouping, respectively. *† indicates that MD from an estimation method is significantly greater than MD calculated from the commonly used method of Yamada II.

Correlations Between Estimated and True Type B Genetic Correlations

Pearson correlation coefficients (calculated based on 300 randomly simulated data samples) between the estimated and true type B genetic correlations for a given pair of environments were generally low for all estimation methods (Table 5-5). Considerable

differences existed, however, among estimation methods. For the constrained multivariate methods, the grouping containing only two environments yielded higher Pearson correlation coefficient between the true and estimated type B genetic correlations than the grouping that included all 4 environment (Table 5-5). Among all estimation methods, the correlation coefficients were highest for the univariate GCA approach, which was followed by Yamada

Table 5-5. Pearson correlation coefficients between estimates of type B genetic correlations and their underlying true values for different estimation methods using simulated half-sib families tested in a randomized complete block experimental design with single-tree plots.

Mortality	Estimation Method	Environment Pairs					
		1-2	1-3	1-4	2-3	2-4	3-4
0%	MTDFREML 1	0.264	0.278	0.279	0.228	0.399	0.207
	MTDFREML 2	0.314	0.336	0.319	0.336	0.424	0.361
	ASREML 1	0.282	0.292	0.254	0.273	0.358	0.272
	ASREML 2	0.252	0.321	0.319	0.193	0.236	0.271
	GCA-approach	0.397	0.405	0.381	0.411	0.454	0.363
	Yamada I	0.311	0.302	0.354	0.354	0.407	0.347
	Yamada II	0.313	0.296	0.360	0.360	0.424	0.365
	Burdon	0.397	0.405	0.381	0.411	0.454	0.363
30%	MTDFREML 1	0.128	0.236	0.162	0.219	0.233	0.075
	MTDFREML 2	0.204	0.290	0.242	0.367	0.269	0.203
	ASREML 1	0.160	0.234	0.082	0.262	0.256	0.165
	ASREML 2	0.181	0.260	0.100	0.257	0.214	0.141
	GCA-approach	0.298	0.329	0.336	0.339	0.360	0.235
	Yamada I	0.205	0.254	0.192	0.280	0.289	0.197
	Yamada II	0.226	0.268	0.270	0.303	0.332	0.247
	Burdon	0.237	0.181	0.162	0.197	0.046	0.067

True genetic parameters for different environments are given in Table 5-1. Out-of-bound estimates were accepted with their original values. † MTDFREML (or ASREML) 1 and 2 refer to the 4-environment and 2-environment grouping, respectively.

II and then multivariate method MTDFREML with 2-environment grouping. Burdon's method was again equal to the GCA approach when data were balanced within an environment, but inferior to the GCA-approach when data were unbalanced within an environment. Results forYamada I and ASREML were between those of the GCA approach and Burdon's method.

Distribution of Estimates

For a given true underlying type B genetic correlation, various estimates were obtained from random data samples due to sampling errors. The scatter plots of estimated type B genetic correlations against their true values were affected by both the estimation methods and the true genetic parameters. Multivariate method MTDFREML and univariate method of Yamada II (Eq.5-5) constrained estimates of type B genetic correlations not greater than 1 and, consequently, skewed the distribution of estimates when true type B genetic correlation was close to 1 (Figures 5-1 & 5-2). Multivariate method ASREML and all other univariate methods, on the other hand, allowed for out-of-bound estimates which made the distribution of estimates appear nearly symmetric against the true values. For unconstrained univariate and multivariate methods, however, some very large estimates were also produced. The frequency of out-of-bound estimates steadily increased as heritabilities in one or two of the environments became smaller (Figure 4-1). For example, between environments 2 and 3, which had heritability 0.3 and 0.2, respectively, there were only 21 (out of 300) estimates greater than 1.0. In contrast, between environments 1 and 4,which had heritabilities of 0.4 and 0.1, respectively, more than 60 (out of 300) estimates were greater than 1.0, although the designed true parameters for both environment pairs were 0.7.

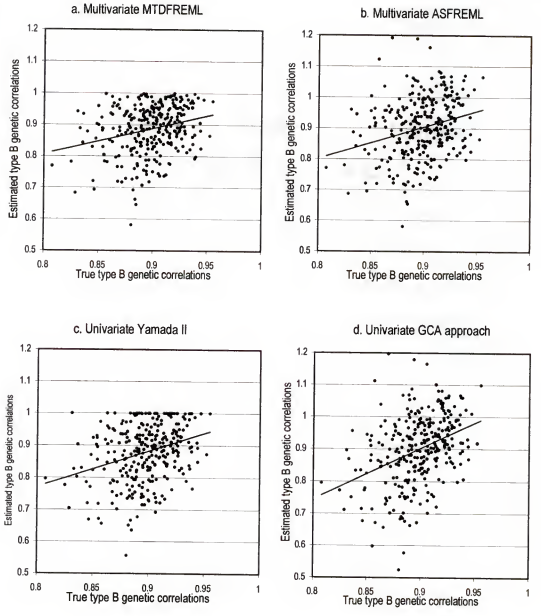

Figure 5-1. Scatter plots of estimates of type B genetic correlations from multivariate and univariate methods against true type B genetic correlations for 300 random samples. True parameters are: h_1^2=0.4, h_2^2=0.3, and r_B=0.9. Experimental designs in both environments are randomized complete block with one tree per family per block. In each environment, there are 20 blocks and 90 half-sib families.

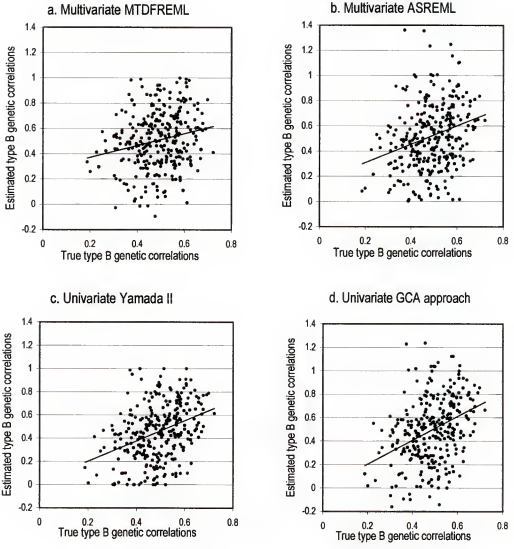

Figure 5-2. Scatter plots of estimates of type B genetic correlations from multivariate and univariate methods against the true type B genetic correlations for 300 random samples. True parameters are: h_1^2=0.2, h_2^2=0.1, and r_B=0.5. Experimental designs in both environments are randomized complete block with one tree per family per block. In each environment, there are 20 blocks and 90 half-sib families.

Discussion

Results of simulations in this study demonstrated that estimates of type B genetic correlations using multivariate methods were empirically as unbiased as the best results from univariate methods (Table 5-3) for both balanced and balanced data with heterogeneous genetic and error variances. Although a tendency of slightly downward bias was detected for the constrained multivariate method MTDFREML, the magnitudes of such biases were very small, hence negligible. This tendency of downward bias was possibly caused by constraining estimates within theoretical parameter space, some of which may otherwise be out-of-bounds. The small magnitudes of such downward biases were probably due to: (1) the relatively small proportion of estimates of type B genetic correlations which would be out-of-bounds between environments having higher heritabilities and, consequently, causing little changes to the mean of estimates when they were constrained within parameter space; and (2) potential compensation by upward biases of estimates of type B genetic correlations for pairs of environments having low heritabilities. For traits of low heritability, there is a high probability that the estimates of genetic variance would be zero or negative (Hill and Thompson 1978). The practice of setting estimates of type B genetic correlations to zero for data samples with zero or negative estimates of genetic variances are likely to yield upwardly biased estimates of type B genetic correlation in prolonged use.

The empirically unbiased estimates of type B genetic correlations from the univariate methods of Yamada I and II were likely due to the specific data structures simulated in this study. Previous numerical studies (Fernando et al. 1984; Dutilleul and Carrière 1998; Chapter 4 of this dissertation) indicated that estimates of type B genetic correlations from

Yamada's methods were subject to bias when heterogeneity of variances was severe and data were highly unbalanced among environments in terms of their relative sizes. In this simulation study, data samples had about the same sizes across all environments, although the genetic and environmental variances were heterogeneous. As a result, biases from Yamada's methods were less severe or negligible.

Because each estimation method was used to analyze the same data sets, differences among estimation methods in MD reflected their differential estimation precisions. The smaller MDs (Table 5-4) obtained from constrained estimation methods (MTDFREML and Yamada II) than from unconstrained methods (ASREML, Yamada I, Burdon, and the GCA approach) were expected because, by theory, the true values of type B genetic correlations cannot be located outside the parameter space. Therefore, higher estimation precision can simply be achieved by restraining estimates of genetic correlations from being out of bounds, which consequently narrows the confidence interval of estimates even though such confidence intervals may not be symmetric.

The smaller MD for a multivariate method MTDFREML which used data from all four environments to estimate the pairwise type B genetic correlation compared to the two-environment grouping may possibly be due to the more stable estimates of genetic variances from the four-environment grouping than from the two-environment grouping. Evidence supporting this reasoning is that the standard deviation of estimates of genetic variance among the 300 random data samples for a given environment was slightly smaller for the four-environment grouping than for the two-environment grouping. This may suggest that, for multivariate methods, additional information in a larger grouping system helped improve

the quality of estimates of genetic variance components so that the sampling error for type B genetic correlation was smaller.

The magnitudes of MDs were surprisingly large for estimates of type B genetic correlations between pairs of environments with traits of low heritabilities. In the simulation, the MDs increased substantially from balanced data with 20 single-tree blocks to unbalanced data with an average of 14 single-tree blocks. With such large sampling errors, the biases of the estimates may become less meaningful. To have reliable estimates of type B genetic correlations a relatively large number of replications and families in the field experimental designs may be required.

Besides the properties of unbiasedness and precision, correlation between estimated and true type B genetic correlation could be important because it reflects the response of estimates to changes on the underlying true values. In this study, the generally low correlation between the estimated and true type B genetic correlation was attributable to the large sampling errors of type B genetic correlations as indicated by the range of estimates for a given true value (Figure 5-2). For the constrained multivariate method, a smaller system containing only two environments seemed to be more desirable in type B genetic correlation estimation than a larger system containing more environments. Interferences by information from unconcerned environments may have reduced the correspondence of estimates of type B genetic correlation to their true values. This trend was clearly demonstrated by MTDFREML with the higher correlation coefficients from the two-environment system than from the four-environment system for various genetic backgrounds. The two-environment system had yielded larger MDs, however, as compared with the four-environment system.

The reason for such results was not clear, probably due to differences between the two grouping systems in variance component estimation.

While constraining estimates of type B genetic correlation within the theoretical parameter space have increased the overall estimation precision, it also skewed the distribution of estimates and caused downward biases. For example, scatter plots of estimates of type B genetic correlations against their true values indicated that for the unconstrained methods, multivariate ASREML and univariate GCA approach, the estimates were nearly symmetrically distributed for a given true value (Figure 5-1b and 5-1d). In contrast, for constrained multivariate method MTDFREML and univariate method Yamada II, the distribution of estimates (Figure 5-1a and 5-1c) was not symmetric against the true values due to the limitation of theoretical boundary, which had definitely altered the distributional pattern of estimates.

The choice of constrained or unconstrained estimates of type B genetic correlations may be objective specific. For example, in theoretical studies of the distribution and sampling errors of type B genetic correlation, unconstrained estimates may be more desirable to show the original distributional pattern of estimates so that the potential confidence intervals of estimates can be investigated in an unbiased manner. In practical genetic data analysis, however, constrained estimates of type B genetic correlations may be easier to interpret and more reasonable to apply when they are involved in indirect selections. For a given set of values of heritabilities, phenotypic variance and selection intensities, genetic response from indirect selection is theoretically less or equal to the gain from direct selection (Falconer 1989). If estimates of genetic correlations greater than 1 were used, however, the

above theoretical rule would be violated, yielding more predicted gains from indirect selection than from direct selection.

Although univariate methods, such as the GCA approach, can achieve estimates of type B genetic correlations as unbiased as those from unconstrained multivariate methods for balanced and unbalanced data, it may be practically more convenient and efficient to use multivariate methods if such computer softwares are available. In addition to the property of constrained estimates with empirical unbiasedness and higher precision, multivariate methods offer additional advantages for data having multiple generations or genetic relatedness that cannot currently be accounted for by univariate methods. These benefits would include: (1) multivariate methods can make use of pedigree information so that type B genetic correlations can be estimated between two environments that have only indirect genetic connectedness, (2) genetic relatedness among genetic groups within and between environments can be accounted for so that the assumption of independence among genetic groups with univariate methods can be relaxed, and (3) there is greater flexibility to include data from mixtures of mating designs and different generations. As tree improvement programs progress into advanced generations and data structures become more complicated (White 1996), multivariate methods could be more appropriate in estimating type B genetic correlations.

Conclusion

Although some univariate methods can yield unbiased estimates of type B genetic correlation for unbalanced data with heterogeneous variances, advantages associated with multivariate methods make them a viable option in the estimation of type B genetic

correlations. Estimates of type B genetic correlations from multivariate methods are empirically unbiased for unbalanced data with heterogeneous variances. Constraining estimates within theoretical parameter space improves estimation precision and practical application. Using more environments in an analytical system with multivariate methods not only increases computational efficiency, but may also enhance the quality of estimates of genetic variances, resulting in smaller sampling errors of the estimates of type B genetic correlations.

CHAPTER 6
CONCLUSIONS

For tree improvement programs, proper and effective analyses of data from genetic tests are crucial for obtaining accurate information about genetic parameters and for achieving maximum genetic gains by correctly predicting breeding values, ranking candidates and making selections. The application of mixed model methods in forest genetic data analysis greatly enhances the effectiveness of handling unbalanced data and improves the properties of predicted breeding values. The use of incomplete mixed linear models with respect to forest genetic experiment helps reduce computational demands for large data sets, but may adversely affect the quality of data analysis by producing considerable biases. The relative consequences of biases produced by incomplete mixed linear models are especially severe for traits under weak additive genetic control but with strong influence of dominance and genotype x environment (G x E) interactions.

For mixed linear models excluding genotype x environment interaction, the magnitude of biases for estimated heritabilities and predicted breeding values are linearly related to the number of testing environments in which progeny tests are arranged. The fewer the number of environments involved, the larger the potential biases. Single-site data analysis represents the extreme example in which the estimates of additive genetic variance components are upwardly biased by the entire G x E interaction variance components.

If dominance effects are excluded by mixed linear models in the analysis of full-sib data, biases to the estimates of heritability and predicted genetic gains are linearly proportional to the numbers of crosses in which each parent is involved. The fewer the number of crosses per parent, the larger the potential biases. The commonly used circular mating design generally has fewer crosses per parent than other mating designs such as the half-diallel, therefore, potentially larger biases would be expected for circular mating design if dominance effects are strong and not included in analytical mixed models.

The purely additive genetic model, which is often used in animal genetic data analysis, ignores both dominance effects and G x E interactions and is often unsuitable in forest genetic data analysis for several reasons: (1) economically important traits of most timber species are usually associated with low heritability; (2) dominance effects and G x E interactions are reasonably strong for many important traits of forest species; and (3) the purely additive model accumulates biases from ignoring both G x E interaction and dominance.

In addition to biases in estimating genetic parameters, incomplete analytical mixed linear models generally give false information about the quality of breeding value and genetic gain prediction. By upwardly biasing the estimates of additive genetic variance components, most incomplete mixed models tend to over predict genetic responses from selection and overstate the reliability of predicted breeding values. Therefore, interpretation of results from incomplete mixed linear models must be conducted with caution, understanding that upward biases may exist.

In circumstances where incomplete mixed linear models need to be used to facilitate data analyses, the best choice would be to ignore the dominance x environment interaction.

Potential biases from this incomplete mixed model are generally downward and several-fold smaller than the biases from incomplete models ignoring other non-additive genetic effects. At the same time, the computational demands are greatly reduced due to the large number of degrees of freedom associated with the dominance x environment interaction.

Despite the potential biases of incomplete linear models in heritability estimation and breeding value prediction, the use of incomplete mixed models does not affect actual genetic gains from selection. This is because the incomplete models yield similar rankings of candidates to the full model, and, therefore, the selected candidates are largely the same candidates whether from full or incomplete models.

The statistical relationship between an incomplete and a full linear model based on balanced data provides useful tools to investigate the potential biases from incomplete linear models. Closed forms can be derived to estimate the biases from a specific incomplete linear models for a given genetic background and field experimental design. Such formulae can be extended to estimate biases for incomplete models when data are moderately unbalanced by using the average design parameters of an experiment.

Type B genetic correlations are frequently estimated in forest genetic data analyses for purposes of studying G x E interactions and making indirect selections. Commonly used univariate methods are derived based on balanced data and often become sub-optimal when data are unbalanced and with heterogeneous variances. A new univariate approach is proposed in this study which is based on predicted parental GCA effects using univariate BLUP program in individual environments. Both theoretical considerations and numerical comparison have shown this is a viable choice to estimate type B genetic correlation when data are highly unbalanced or/and have heterogenous variances among environments.

Empirical properties of estimates from this new GCA approach include unbiasedness, relatively high precision and a better relationship with the true underlying type B genetic correlations.

While univariate estimation methods of type B genetic correlations are exclusively used in forest genetic data analyses due to their small computational demand and availability of computer software, multivariate methods were shown in this study to yield more desirable estimates. By treating observations in different environments as different traits and incorporating them into a closed analytical system, multivariate methods estimate genetic variance and covariance components simultaneously using the REML approach. Constrained estimates of type B genetic correlations from multivariate methods are all within the theoretical parameter space and empirically unbiased. By eliminating out-of-bound estimates of type B genetic correlation, the average precision and accuracy of estimates can be considerably improved due to the fact that true type B genetic correlation cannot be out of parameter space. The higher estimation precision of the constrained multivariate method manifests itself more strikingly when heritabilities of a trait under investigation are low in some environments. Practically, constrained estimates of type B genetic correlations are more reasonable for interpretation and more appropriate for use in estimating genetic gains from indirect selections.

Practical application of multivariate methods in type B genetic correlation estimation should become popular when use-friendly computer softwares are widely available and genetic testing data have complex relatedness structures. Proper data structures and correct determination of analytical mixed linear models are vital to use a multivariate program.

APPENDIX 2-1

EXPECTED SUM OF SQUARES FOR THE EFFECT OF FAMILY X BLOCK (ENVIRONMENT) INTERACTION IN THE INCOMPLETE LINEAR MODEL IGNORING THE EFFECT OF FAMILY X ENVIRONMENT INTERACTION

FromEq. 2-4, page 12,

$$E(SS*_{FxB/E}) = E[n\sum_{i=1}^{t}\sum_{j=1}^{b}\sum_{k=1}^{f}(\overline{y_{ijk.}}-\overline{y_{ij..}}-\overline{y_{..k}}+\overline{y_{...}})^2] = n\sum_{i=1}^{t}\sum_{j=1}^{b}\sum_{k=1}^{f}E[(\overline{y_{ijk.}}-\overline{y_{ij..}}-\overline{y_{..k}}+\overline{y_{...}})^2]$$

$$= n\sum_{i=1}^{t}\sum_{j=1}^{b}\sum_{k=1}^{f}E[(\mu+E_i+B_{ij}+f_k+p_{ijk}+\overline{e_{ijk.}}-\mu-E_i-B_{ij}-\overline{f}_{.}-\overline{p_{ij.}}-\overline{e_{ij..}}$$

$$-\mu-E_{.}-B_{..}-f_k-\overline{p}_{..k}-\overline{e}_{..k}+\mu+E_{.}+B_{..}+\overline{f}_{.}+\overline{p}_{...}-\overline{e}_{...})^2]$$

$$= n\sum_{i=1}^{t}\sum_{j=1}^{b}\sum_{k=1}^{f}E(p_{ijk}+\overline{e_{ijk.}}-\overline{p_{ij.}}-\overline{e_{ij...}}-\overline{p}_{..k}-\overline{e}_{..k}+\overline{p}_{...}+\overline{e}_{...})^2$$

due to $E(p_{ijk}+\overline{e_{ijk.}}-\overline{p_{ij.}}-\overline{e_{ij...}}-\overline{p}_{..k}-\overline{e}_{..k}+\overline{p}_{...}+\overline{e}_{...})=0$

$$= n\sum_{i=1}^{t}\sum_{j=1}^{b}\sum_{k=1}^{f}[V(p_{ijk})+V(\overline{e_{ijk.}})+V(\overline{p_{ij.}})+V(\overline{e_{ij..}})+V(\overline{p}_{..k})+V(\overline{e}_{..k})+V(\overline{p}_{...})+V(\overline{e}_{...})-2Cov(p_{ijk},\overline{p}_{ij.})$$

$$-2Cov(p_{ijk},\overline{p}_{..k})+2Cov(p_{ijk},\overline{p}_{...})+2Cov(\overline{p_{ij.}},\overline{p}_{..k})-2Cov(\overline{p_{ij.}},\overline{p}_{...})-2Cov(\overline{p}_{..k},\overline{p}_{...})-2Cov(\overline{e_{ijk.}},\overline{p_{ij.}})$$

$$-2Cov(\overline{e_{ijk.}},\overline{e}_{..k})+2Cov(e_{ijk.},\overline{e}_{...})+2Cov(\overline{e_{ij..}},\overline{e}_{..k})-2Cov(e_{ij..},\overline{e}_{...})-2Cov(\overline{e}_{..k},\overline{e}_{...})$$

$$= n\sum_{i=1}^{t}\sum_{j=1}^{b}\sum_{k=1}^{f}(\sigma^2_{p*}+\frac{1}{n}\sigma^2_{e*}+\frac{1}{f}\sigma^2_{p*}+\frac{1}{fn}\sigma^2_{e*}+\frac{1}{tb}\sigma^2_{p*}+\frac{1}{tbn}\sigma^2_{e*}+\frac{1}{tbf}\sigma^2_{p*}+\frac{1}{tbfn}\sigma^2_{e*}-\frac{2}{f}\sigma^2_{p*}-\frac{2}{tf}\sigma^2_{p}$$

$$+\frac{2}{tbf}\sigma^2_{p*}+\frac{2}{tbf}\sigma^2_{p*}-\frac{2}{tbf}\sigma^2_{p*}-\frac{2}{fn}\sigma^2_{e*}-\frac{2}{tbn}\sigma^2_{e*}+\frac{2}{tbfn}\sigma^2_{e*}+\frac{2}{tbfn}\sigma^2_{e*}-\frac{2}{tbfn}\sigma^2_{e*}-\frac{2}{tbfn}\sigma^2_{e*}$$

$$= bfnt[(1-\frac{1}{f})(1-\frac{1}{tb})(\sigma^2_{p*}+\frac{1}{n}\sigma^2_{e*})$$

$$= (f-1)(tb-1)(\sigma^2_{e*}+n\sigma^2_{p*}).$$

APPENDIX 2- 2

EXPECTED SUM OF SQUARES FOR THE FAMILY EFFECT IN THE INCOMPLETE LINEAR MODEL IGNORING THE EFFECT OF FAMILY X ENVIRONMENT INTERACTION

From Eq.2-6 page 13, we have

$$E[SS*_{family}]=E[bnt\sum_{k=1}^{f}(\overline{y_{..k}}-\overline{y_{....}})^2]$$

$$=bnt\sum_{k=1}^{f}E[(\mu+\overline{E_{.}}+\overline{B_{..}}+f_k+\overline{p_{..k}}+\overline{e_{..k}}-\mu-\overline{E_{.}}-\overline{B_{..}}-\overline{f_{.}}-\overline{p_{...}}-\overline{e_{....}})^2]$$

$$=bnt\sum_{k=1}^{f}E[(f_k+\overline{p_{..k}}+\overline{e_{..k}}-\overline{f_{.}}-\overline{p_{...}}-\overline{e_{....}})^2]$$

$$=bnt\sum_{k=1}^{f}V(f_k+\overline{p_{..k}}+\overline{e_{..k}}-\overline{f_{.}}-\overline{p_{...}}-\overline{e_{....}}) \qquad \text{due to} \quad E(f_k+\overline{p_{..k}}+\overline{e_{..k}}-\overline{f_{.}}-\overline{p_{...}}-\overline{e_{....}})=0$$

$$=bnt\sum_{k=1}^{f}[V(f_k)+V(\overline{p_{..k}})+V(\overline{e_{..k}})+V(\overline{f_{.}})+V(\overline{p_{...}})+V(\overline{e_{....}})$$

$$=bnt\sum_{k=1}^{f}(\sigma^2_{f*}+\frac{1}{bt}\sigma^2_{p*}+\frac{1}{bnt}\sigma^2_{e*}+\frac{1}{f}\sigma^2_{f*}+\frac{1}{bft}\sigma^2_{p*}+\frac{1}{bfnt}\sigma^2_{e*}-\frac{2}{f}\sigma^2_{f*}-\frac{2}{bft}\sigma^2_{p*}-\frac{2}{bfnt}\sigma^2_{e*}]$$

$$=bfnt[(1-\frac{1}{f})\sigma^2_{f*}+\frac{1}{bt}(1-\frac{1}{f})\sigma^2_{p*}+\frac{1}{bnt}(1-\frac{1}{f})\sigma^2_{e*}]$$

$$=(f-1)(\sigma^2_{e*}+n\sigma^2_{p*}+bnt\sigma^2_{f*}).$$

APPENDIX 4-1
EXPECTED VALUES OF COVARIANCE AND VARIANCES OF ADJUSTED PARENTAL GCA EFFECTS

From Chapter 4 page 69, it is defined that predicted parental GCA effects are $\hat{g}_{xi}$ and $\hat{g}_{yi}$, and their true value be g_{xi} and g_{yi}, respectively, in environments **x** and **y**. Further, the prediction accuracies $\hat{g}_{xi}$ and $\hat{g}_{yi}$ are r_{axi} and r_{ayi}, respectively, in two environments, then

$$E[\hat{\sigma}_{gxy}] = E\left[\frac{1}{m}\sum_{i=1}^{m} Cov_i\left(\frac{\hat{g}_{xi}}{r_{axi}}, \frac{\hat{g}_{yi}}{r_{ayi}}\right) / \overline{r_{axi} r_{ayi}}\right]$$

$$= E\left[\frac{1}{m}\sum_{i=1}^{m} Cov_i\left(\frac{\hat{g}_{xi}}{r_{axi}^2}, \frac{\hat{g}_{yi}}{r_{ayi}^2}\right) \frac{r_{axi} r_{ayi}}{\overline{r_{ax} r_{ay}}}\right]$$

$$= E\left[\frac{1}{m}\sum_{i=1}^{m} Cov_i\left(\frac{\hat{g}_{xi}}{b_{xi}}, \frac{\hat{g}_{yi}}{b_{yi}}\right) \frac{r_{axi} r_{ayi}}{\overline{r_{ax} r_{ay}}}\right]$$

[b_{xi} is regression coefficient and $b_{xi} = r_a^2$, (Mrode 1996]

Because $\hat{g}_{xi} = b_{\bar{x}i}(\bar{x}_i - \hat{\tau}) = b_{\bar{x}i} * (\mu + \bar{b}_{x.} + g_i + \bar{e}_{.} - \hat{\tau}) = b_{\bar{x}i} * (g_{xi} + \bar{e}_{x.})$,

and similarly, $\hat{g}_{yi} = b_{\bar{y}i} * (g_{yi} + \bar{e}_{y.})$,

also, $Cov[(g_{xi} + \bar{e}_{x.}), (g_{yi} + \bar{e}_{y.})] = Cov(g_{xi}, g_{yi})$ (residuals in **x** and **y** are uncorrelated),

thus, by substitution,

$$E\left[\frac{1}{m}\sum_{i=1}^{m} Cov_i\left(\frac{\hat{g}_{xi}}{b_{xi}}, \frac{\hat{g}_{yi}}{b_{yi}}\right) \frac{r_{axi} r_{ayi}}{\overline{r_{ax} r_{ay}}}\right]$$

$$= E\left[\frac{1}{m}\sum_{i=1}^{m} Cov_i(g_{xi}, g_{yi}) \frac{r_{axi} r_{ayi}}{\overline{r_{ax} r_{ay}}}\right]$$

$$= E\left[\frac{1}{m}\sum_{i=1}^{m} \sigma_{gxy} \frac{r_{axi} r_{ayi}}{\overline{r_{ax} r_{ay}}}\right]$$

$$= E\left[\frac{1}{m}\sum_{i=1}^{m}\sigma_{gxy}\frac{r_{axi}\,r_{ayi}}{r_{ax}\,r_{ay}}\right]$$

$$= E\left[\frac{\sigma_{gxy}}{m}\sum_{i=1}^{m}\frac{r_{axi}\,r_{ayi}}{r_{ax}\,r_{ay}}\right]$$

$$= E\left[\sigma_{gxy}\frac{\overline{r_{ax}\,r_{ay}}}{r_{ax}\,r_{ay}}\right] = E[\sigma_{gxy}] = \sigma_{gxy}.$$

$$E\left[Var\left(\frac{\hat{g}_{xi}}{r_{xi}}\right)\right] = E\left[Var\left(\frac{b_{\bar{x}i}\,\overline{x}_i}{r_{xi}}\right)\right] = E\left[Var\left(\frac{b_{\bar{x}i}\,\overline{x}_i}{\sqrt{b_{\bar{x}i}}}\right)\right]$$

$$= E\left[Var\left(\sqrt{b_{\bar{x}i}}\,\overline{x}_i\right)\right] = E\left[b_{\bar{x}i}\,Var\left(\overline{x}_i\right)\right]$$

$$= E[\frac{\sigma_{gx}^2}{(\sigma_{gx}^2+\frac{\sigma_{ex}^2}{n_{xi}})}(\sigma_{gx}^2+\frac{\sigma_{ex}^2}{n_{xi}})] = E[\sigma_{gx}^2] = \sigma_{gx}^2.$$

Similarly,

$$E\left[Var\left(\frac{\hat{g}_{yi}}{r_{ayi}}\right)\right] = E\left[Var\left(\frac{b_{\bar{y}i}\,\overline{y}_i}{r_{ayi}}\right)\right] = E\left[Var\left(\frac{b_{\bar{y}i}\,\overline{y}_i}{\sqrt{b_{\bar{y}i}}}\right)\right]$$

$$= E\left[Var\left(\sqrt{b_{\bar{y}i}}\,\overline{y}_i\right)\right] = E\left[b_{\bar{y}i}\,Var\left(\overline{y}_i\right)\right]$$

$$= E[\frac{\sigma_{gy}^2}{(\sigma_{gy}^2+\frac{\sigma_{ey}^2}{n_{yi}})}(\sigma_{gy}^2+\frac{\sigma_{ey}^2}{n_{yi}})] = E[\sigma_{gy}^2] = \sigma_{gy}^2.$$

REFERENCES

Adams, W.T., White, T.L., Hodge, G.R., and Powell, G.L. 1994. Genetic parameters for bole volume in longleaf pine: large sample estimates and influences of test characteristics. Silvae Genet. 43:357-66.

Araújo, J.A., Sousa, R., Lemos, L., and Borralho, N.M.G. 1996. Estimates of genetic parameters and prediction of breeding values for growth in *Eucalyptus globulus* combining clonal and full-sib progeny information. Silvae Genet. 45:223-26.

Boldman, K.G., Kriese, L.A., Van Vleck, L.D., Van Tassell, C.P., and Kachman, S.D. 1995. A manual for use of MTDFREML. A set of programs to obtain estimates of variances and covariances [draft]. U.S. Department of Agriculture, Agricultural Research Service.

Bongarten, B.C., and Hanover J.W. 1986. Genetic parameters of blue spruce (*Picea pungens*) at two locations in Michigan. Silvae Genet. 35:106-12.

Borralho, N.M.G., and Wilson, Ph.J. 1994. Inheritance of initial survival and rooting ability in Eucalyptus Globulus Labill. stem cuttings. Silvae Genet. 43:238-42.

Borralho, N.M.G., and Dutkowski, G.W. 1998. comparison of rolling front and discrete generation breeding strategies for trees. Can. J. For. Res. 28:987-93.

Burdon, R.D. 1977. Genetic correlation as a concept for studying genotype-environment interaction in forest tree breeding. Silvae Genet. 26:168-75.

Burdon, R.D., and Van Buijtenen, J.P. 1990. Expected efficiencies of mating designs for reselection of parents. Can. J. For. Res. 20:1664-71.

Campbell, R.K., Echols, R.M., and Stonecypher, R.W. 1986. Genetic variances and interactions in 9-year-old Douglas-fir grown at narrow spacings. Silvae Genet. 35:24-32.

Cotterill, P.P., and James, J.W. 1982. Number of offspring and plot sizes required for progeny testing. Silvae Genet. 33:203-209.

Cooper, M., and Delacy, I.H. 1994. Relationship among analytical methods used to study genotypic variation and genotype-by-environment interaction in plant breeding multi-environment experiments. Theo. Appl. Genet. 88:561-72.

de Boer, I.J.M., and van Arendonk, J.A.M. 1992. Prediction of additive and dominance effects in selected or unselected populations with inbreeding. Theor. Appl. Genet. 84:451-59.

de Boer, I.J.M., and Hoeschele, I. 1993. Genetic evaluation methods for populations with dominance and inbreeding. Theor. Appl. Genet. 86:245-58.

Dickerson, G.E. 1962. Implications of genetic environment interaction in animal breeding. Animal Production. 4:47-64.

Dieters, M.J.J. 1996. Genetic parameters for slash pine (*Pinus elliottii*) grown in south-east Queensland, Australia: growth, stem straightness and crown defects. Forest Genetics. 3(1): 27-36.

Dieters, M.J., White, T.L., and Hodge, G.R. 1995. Genetic parameter estimates for volume from full-sib tests of slash pine (*Pinus elliottii*). Can. J. For. Res. 25:1397-408.

Dutilleul, P., and Carrière, Y. 1998. Among-environment heteroscedasticity and the estimation and testing of genetic correlation. Heredity 80:403-13.

Ericsson, T., and Danell, Ö. 1995. Genetic evaluation, multiple-trait selection criteria and genetic thinning of *Pinus contorta* var. *latiforlia* seed orchards in Sweden. Scand. J. For. Res. 10:313-25.

Falconer, D.S. 1952. The problem of environment and selection. Amer. Nat. 86:293-98.

Falconer, D.S. 1981. Introduction to quantitative genetics. Longman & Co., New York, NY.

Fernando, R.L., Knights, S.A., and Gianola, D. 1984. On a method of estimating the genetic correlation between characters measured in different experimental units. Theor. Appl. Genet. 67:175-78.

Foster, G.S., and Bridgewater, F.E. 1986. Genetic analysis of fifth-year data from a seventeen parent partial diallel of loblolly pine. Silvae Genet. 35:18-122.

Giertych, M., and Van De Sype, H. 1990. Consequences of reducing a full model of variance analysis in tree breeding experiments. Ann. Sci. For. 47:17-29.

Giesbrecht, G.S. 1983. An efficient procedure for computing MINQUE of variance components and generalized least squares estimates of fixed effects. Commun. Stat. Theory Meth. 12:2169-77.

Gilmour, A.R., Thompson, R, Cullis, B.R. and Welham, S.J. 1997. ASREML, NSW Agriculture, ORANGE, 2800, Australia.

Hardner, C., Borralho, N., Tier, B., Miller, S., and Goddard, M. 1996. Accounting for dominance and inbreeding in genetic evaluations using individual tree mixed models. In tree improvement for sustainable tropic forestry. Dieters M.J., Matheson, A.C., Nikles, D.G., Harwood, C.E., Walker, S.M. (eds) Prox. QFRI-IUFRO Conf, Caloundra, Queensland, Australia. 143-47.

Henderson, C.R. 1953.Estimation of variance and covariance components. Biometrics. 9:226-22.

Henderson, C.R. 1973. Sire evaluation and genetic trends. *In* Animal Breeding and Genetics Symposium in Honor of J. Lush, Champaign, Ill: Animal Sci. Assoc. Amer. P10-41.

Henderson, C.R. 1977. Prediction of future records. In Proc. Internat. Conf. Quan. Genet. E. Pollack, O. Kempthorne and T. Bailey (eds.) The Iowa State University Press, Ames, IA. 615-38.

Henderson, C.R. 1984. Application of linear models in animal breeding. University of Guelph, Guelph, Ontario, Canada.

Henderson, C.R. 1985. Best linear unbiased prediction of nonadditive genetic merits in non-inbred populations. J. Anim. Sci. 60:111-17.

Hill, W.G., and Thompson, R. 1978. Probability of non-positive define between-group or genetic covariance matrices. Biometrics. 34:429-39.

Hodge,G.R., and Purnell, R.C. 1993. Genetic parameter estimates for wood density, transition age and radial growth in slash pine. Can. J. For. Res. 23:1881-1891.

Hodge, G.R., and White, T.L. 1992. Genetic parameter estimates for growth traits at different ages in slash pine and some implications for breeding. Silvae Genet. 41:252-62.

Huber, D.A., White, T.L., and Hodge, G.R. 1992. The efficiency of half-sib, half-diallel and circular mating designs in the estimation of genetic parameters in forestry: a simulation. Forest Science. 38:757-76.

Huber, D.A. 1993. Optimal mating designs and optimal techniques for analysis of quantitative traits in forest genetics. Ph.D Dissertation, Univ. Florida, Gainesville, FL.

Huber, D.A., White, T.L., and Hodge, G.R. 1994. Variance component estimation techniques compared for two mating designs with forest genetic architecture through computer simulation. Theor. Appl. Genet. 88:236-42.

Itoh, Y. and Yamada, Y. 1990. Relationship between genotype x environment interaction and genetic correlation of the same trait measured in different environments. Theor. Appl. Genet. 80:11-16.

Jarvis, S.F., Borralho, N.M.G., and Potts, B.M. 1995. Implementation of a multivariate BLUP model for genetic evaluation of *Eucalyptus globulus* in Australia. In Proc. CRC-IUFRO Conf. Eucalypt Plantations: Improving Fibre Yield and Quality. Hobart, Tasmania, Australia. 212-16.

Jiang, I.B.J. 1985. Early testing in forest tree breeding: a review. In Forest tree improvement No.20 1987. Proc. from a meeting on early testing, juvenile-mature correlations and accelerated generation turnover. Horsholm, Denmark. 45-78.

Johnson, G.R. 1997. Site-to-site genetic correlation and their implications on breeding zone size and optimal number of progeny test sites for coast Douglas-fir. Silvae Genet. 46:280-85.

Johnson, G.R., and Burdon, R.D. 1990. Family-site interaction in *Pinus radiata*: implications for progeny testing strategy and regionalized breeding in New Zealand. Silvae Genet. 39:55-62.

Koots, K.R., and Gibson, J.P. 1996. Realized sampling variances of estimates of genetic parameters and the difference between genetic and phenotypic correlations. Genetics 143:1409-16.

Lambeth, C.C. and Gladstone, W.T. 1983. Statistical efficiency of row and noncontiguous family plot in genetic test of loblolly pine. Silvae Genet. 32:24-28.

Li, B., McKeand, S.E., and Weir, R.J. 1996. Genetic parameter estimates and selection efficiency for the loblolly pine breeding in the southeastern U.S.. In tree improvement for sustainable tropic forestry. Dieters, M.J., Matheson, A.C,, Nikles, D.G., Harwood C.E., Walker, S.M. (eds). Prox. QFRI-IUFRO Conf, Caloundra, Queensland, Australia. 164-68.

Loo-Dinkins, J.A. and Tauer, C.G. 1987. Statistical efficiency of six progeny test field designs on three loblolly pine site types. Can. J. For. Res. 17:1066-70.

Loo-Dinkins, J.A., Tauer, C.G., and Lambeth, C.C. 1990. Selection system efficiencies for computer simulated progeny test field designs in loblolly pine. Theor. Appl. Genet. 79:89-96.

Matheson, A.C. 1989. Statistical methods and problems in testing large numbers of genotype across sites. In Breeding tropic trees: population structure and genetic improvement strategies in clonal and seedling forestry. Oxford For. Ins. Oxford, UK. 93-105.

Montgomery, D.C. 1991. Design and analysis of experiment. John Wiley & Sons, New York.

Mrode, R.A. 1996. Linear models for the prediction of animal breeding values. CAB International, Wallingford, Oxford, UK.

Namkoong, G., and Roberds, J.H. 1974. Choosing mating designs to efficiently estimate genetic variance components for trees. Silvae Genet. 23:43-53.

Patterson, H.D., and Thompson, R. 1971. Recovery of interblock information when block sizes are unequal. Biometrika 58:545-54.

Pswarayi, I.Z., Barnes, R.D., Birks, J.S. and Kanawski, P.J. 1997. Genotype-environment-interaction in a population of *Pinus elliottii* Engelm. var. *elliottii*. Silvae Genet. 46:35-40.

Quinton, V.M., and Smith C. 1997. An empirical check on best linear unbiased prediction genetic evaluation using pig field recording data. Can. J. Anim. Sci. 77:211-16.

Robertson, A. 1959. The sampling variance of the genetic correlation coefficient. Biometrics 15: 469-85.

SAS Institute, 1990. SAS language, SAS Institute Inc. Cary, NC 27513.

SAS Institute, 1993. SAS/STAT User's Guide, Vol. 2. GLM-VARCOMP. Version 6, 4th edition. SAS Institute Inc. Cary, NC 27513.

Schaeffer, L.R., and Wilton, J.W. 1978. Simultaneous estimation of variance and covariance components for multi-trait mixed model equations. Biometrics 34:199-208.

Searle, S.R., Casella, G., and McCulloch, C.E. 1992. Variance Components. John Wiley & Sons, Inc., New York.

Singh, M., Ceccarelli, S., and Grando,S. 1997. Precision of genetic correlation estimation from variety trials conducted in incomplete block designs. Theor. Appl. Genet. 95:1044-48.

Surles, S.E. 1993. Early Selection for volume growth in slash pine. Ph.D dissertation, University of Florida, Gainesville.

Uimari, P., and Mäntysaari, E.A.1993. Repeatability and bias of estimated breeding values for dairy bulls and bull dam calculated from animal model evaluation. Anim. Prod. 57:175-82.

Van Vleck, L.D. 1993. Selection index and introduction to mixed model methods. Boca Raton: CRC Press.

Varona, L., Miaztal, I. Bertrand, J.K., and Lawlor T.J. 1997. Effect of full sibs on additive breeding values under the dominance model for stature in United States Holsteins. J. Diary Sci. 81:1126-35.

Wei, M., and van der Werf, J.H.J. 1993. Animal model estimation of additive and dominance variances in egg production traits of poultry. J. Anim. Sci. 71:57-65.

Wei, X., and Borralho, N.M.G. 1998. Use of individual tree mixed models to account for mortality and selective thinning when estimating base population genetic parameters. For. Sci. 44:246-253.

White, T.L. 1987. A conceptual framework for tree improvement programs. New Forests 4:325-342.

White, T.L. 1996. Genetic parameter estimates and breeding value prediction: issues and implications in tree improvement programs. In Tree improvement for sustainable tropic forestry. Dieters, M.J., Matheson, A.C., Nikles D.G., Harwood, C.E., Walker, S.M. (eds). Prox. QFRI-IUFRO Conf, Caloundra, Queensland, Australia.110-17.

White, T.L., and Hodge, G.R. 1989. Predicting breeding values with application in forest tree improvement. Dordrecht: Kluwer Academic Publishers.

White, T.L., Hodge, G.R., and Powell, G.L. 1993. An advanced-generation tree improvement plan for slash pine in the southeastern United States. Silvae Genet. 42:359-71.

Woolaston, R.R., Kanawski,P.J., and D.G. Nikles,D.G. 1991. Genotype-environment interaction in Pinus caribaea var. hondurensis in Queensland, Australia. II. Family x site interaction. Silvae Genet. 40:228-32.

Wu, X. M. 1993. Early selection in Lodgepole pine. Ph.D. thesis, University of Alberta. Edmonton, Alberta. Canada.

Yamada, Y. 1962. Genotype by environment interaction and genetic correlation of the same trait under different environments. Jap. J. Genet. 37:498-509.

Yeh, F.C., and Heaman, J.C. 1987. Estimating genetic parameters of height growth in seven-year old coastal Douglas-Fir from disconnected diallels. For. Sci. 33:946-57.

Zobel, B.J., and Talbet, J. 1984. Applied Forest Tree Improvement. Wiley and Sons, Inc., New York.

BIOGRAPHICAL SKETCH

Pengxin Lu was born August 4, 1960 in Shaanxi province, People's Republic of China. He was admitted to the Northwestern College of Forestry when he graduated from the Yongshou County High School in 1978. He received his Bachelor of Agriculture degree in 1982. From 1982 to 1992 he was employed by the Chinese Academy of Forestry (CAF) in Beijing as a research scientist in silviculture, tree introduction and genetic improvement. He pursued his Master of Science degree in forest genetics and tree improvement with Dr. Francis C. Yeh at the University of Alberta in Edmonton, Canada, from 1992 to 1994. He entered the Doctor of Philosophy program at the University of Florida in January 1995 after receiving his M.Sc. degree.

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Timothy L. White, Chairman
Professor of Forest Resources
and Conservation

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Dudley A. Huber
Associate in Forest Resources
and Conservation

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Ramon C. Littell
Professor of Statistics

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Donald L. Rockwood
Professor of Forest Resources
and Conservation

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

David S. Wofford
Professor of Agronomy

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

May, 1999

Director, Forest Resources and Conservation

Dean, Graduate School